A Statistical Approach to Topological Data Analysis

Bertrand Michel

MÉMOIRE D'HABILITATION À DIRIGER DES RECHERCHES

Université Pierre et Marie Curie

Laboratoire de Statistique Théorique et Appliquée

Bertrand MICHEL

A Statistical Approach to Topological Data Analysis

soutenue le 24 novembre 2015

devant le jury composé de :

M. Gérard Biau	Université Pierre et Marie Curie	Examinateur
M. Benoît Cadre	ENS Rennes	Rapporteur
M. Frédéric Chazal	INRIA Saclay	Examinateur
M. Albert Cohen	Université Pierre et Marie Curie	Président du jury
M. Wolfgang Polonik	University of California at Davis	Rapporteur
M. Pascal Massart	Université Paris Sud	Examinateur

et au vu du rapport également écrit par :

M. Shmuel Weinberger	University of Chicago	Rapporteur

Remerciements

Je souhaite tout d'abord remercier ici chaleureusement Benoît Cadre, Wolfgang Polonik et Shmuel Weinberger d'avoir accepté de rapporter mon mémoire d'habilitation. Wolfgang, I am very honored that you have accepted to review my habilitation thesis and that you came from Davis for my defense. Shmuel, I am also very honored that you have accepted to review my statistical works about topological data analysis.

Le travail présenté dans ce mémoire est le fruit de collaborations et d'interactions avec de nombreuses personnes que je souhaite vivement remercier.

Je souhaite en premier lieu remercier Pascal Massart et Thomas Duquesne pour m'avoir initié pendant ma thèse au métier d'enseignant-chercheur. Pascal, le soutien que tu as apporté par la suite à ce "mariage" de la statistique et de l'inférence géométrique a été très important pour moi. Thomas, un jour peut-être publierons-nous ce vieux papier qui traîne dans nos placards, mais ce n'est pas le bon moment, le baril est au plus bas.

Je tiens ensuite à remercier très vivement Frédéric Chazal, qui en dépit de mon passé peu recommandable de "statisticien pétrolier", m'a accueilli en post-doc à l'INRIA dans l'équipe Geometrica pour travailler sur ces problématiques mêlant statistique, géométrie algorithmique et analyse des données. Frédéric, je suis très heureux de collaborer avec toi sur ces jolies questions. Je te remercie aussi pour la place que tu m'as faite au sein de l'équipe et j'espère que je pourrai continuer à participer activement aux activités de l'équipe DataShape qui devrait prochainement remplacer Geometrica.

J'ai eu la chance de co-encadrer la thèse de Baptiste Gregorutti avec Philippe Saint-Pierre et Gérard Biau. Merci à toi Baptiste pour avoir tenu bon dans cette co-direction à trois, je suis pour ma part ressorti très heureux de cette première expérience d'encadrement.

Je souhaite aussi adresser un remerciement à l'ensemble de mes co-auteurs qui ont bien évidemment contribué de façon directe à cette habilitation. Je remercie tout d'abord Cathy Maugis avec qui j'ai activement collaboré pendant et après mes années de thèse. Je veux aussi remercier Jérôme Dedecker avec qui j'ai eu beaucoup de plaisir à découvrir des thématiques statistiques, notamment la déconvolution, qui m'étaient inconnues à l'issue de ma thèse. Je remercie aussi mes co-auteurs Jean-Patrick Baudry, Vincent Brault, Claire Caillerie, Frédéric Chazal, Aurélie Fischer, Stéphane Gaïffas, Marc Glisse, Baptiste Gregorutti, Catherine Labruère, Pascal Massart, Philippe Saint-Pierre. J'ai enfin eu la chance de collaborer ces dernières années avec des membres de l'équipe du département de statistique de CMU : Brittany Fasy, Jisu Kim, Fabrizio Lecci, Alessandro Rinaldo et Larry Wasserman.

Depuis la fin de ma thèse et mon recrutement au LSTA, j'ai partagé mes activités de recherche entre la laboratoire de Statistique Théorique et Appliquée de l'UPMC et l'équipe Geometrica de l'INRIA. Je salue et remercie l'ensemble des membres des deux équipes, permanents et doctorants. Cette bi-localisation a été une grande chance pour moi. Je remercie tout particulièrement Gérard Biau non seulement pour les conditions de travail qui m'ont été offertes au LSTA, mais aussi pour m'avoir toujours encouragé à développer ma propre thématique de recherche avec Geometrica, hors des murs du LSTA. Je souhaite aussi saluer ici les collègues du LPMA, les portes coupe-feu ne nous ont heureusement jamais empêchés de prendre des cafés ensemble.

Mes ultimes remerciements vont à ma compagne Amélie pour son soutien infaillible, mais aussi à mes deux garçons Lucien et Gaston qui savent mieux que quiconque ramener leur père à la réalité et aux joies du quotidien, surtout à trois heures du matin.

Abstract

Until very recently, topological data analysis and topological inference methods mostly relied on deterministic approaches. The major part of this habilitation thesis presents a statistical approach to such topological methods. We first develop model selection tools for selecting simplicial complexes in a given filtration. Next, we study the estimation of persistent homology on metric spaces. We also study a robust version of topological data analysis. Related to this last topic, we also investigate the problem of Wasserstein deconvolution. The second part of the habilitation thesis gathers our contributions in other fields of statistics, including a model selection method for Gaussian mixtures, an implementation of the slope heuristic for calibrating penalties, and a study of Breiman's permutation importance measure in the context of random forests.

Keys words: topological data analysis, topological inference, persistent homology, non parametric statistics, model selection, bootstrap, deconvolution, Wasserstein metrics, mixture models, slope heuristics, random forests, permutation importance measure.

Résumé

Jusqu'à très récemment, l'analyse topologique des données ainsi que les méthodes d'inférence topologique ont principalement été développées dans une perspective déterministe. La partie principale de cette thèse d'habilitation traite d'une approche statistique de ces méthodes topologiques. Nous proposons tout d'abord des outils de sélection de modèle pour choisir un complexe simplicial dans une filtration donnée. Nous étudions ensuite le problème de l'estimation de l'homologie persistante. Nous considérons aussi une version robuste de l'analyse topologique des données ainsi que le problème de la déconvolution Wasserstein, ces deux questions étant en fait reliées. La seconde partie de cette thèse d'habilitation rassemble nos contributions dans d'autres domaines de la statistique. Nous y présentons des résultats de sélection de modèle pour des modèles de mélange Gaussien, une implémentation efficace de l'heuristique de pente ainsi qu'une étude de la mesure d'importance de Breiman.

Mots clés : analyse topologique des données, inférence topologique, homologie persistante, statistique non paramétrique, sélection de modèles, bootstrap, déconvolution, métriques Wasserstein, modèles de mélange, heuristique de pente, forêts aléatoires, mesure d'importance par permutation.

Foreword

With the emergence of distance-based approaches and persistent topology, geometric inference and computational topology have recently undergone significant developments. New mathematically well-founded theories have given birth to the field of topological data analysis. Our research mainly involves the statistical analysis of these methods but also includes development of statistical tools and methods in this field. The first and main part of this document presents our contributions to this topic.

The first chapter is an introductory one about topological data analysis and topological inference. It ends by explaining the reasons for a statistical approach to these problems. Chapter 2 is about a model selection method for selecting a simplicial complex in a filtration. Chapter 3 presents our statistical results about persistent homology inference. Chapter 4 looks at a robust version of topological data analysis based on a notion of distance to measure. We also present in Chapter 4 our results about the Wasserstein deconvolution problem, which is related to the problem of robust topological inference.

The second part of the habilitation thesis gathers our contributions on other topics in statistics. Chapter 5 presents our contributions in model selection in the context of clustering with Gaussian mixture models. Chapter 6 details the implementation of the slope heuristic method. Lastly, in Chapter 7 examines feature selection in the context of random forests.

The organization of this document into two parts might lead one to believe that these two parts are totally independent, but this not the case. Indeed, Chapter 5 is about clustering, but topological data analysis is also concerned with this problem of "connectivity" in data. Moreover, Chapter 2 (selection of simplicial complexes) and Chapter 5 (selection of Gaussian mixture models) both relie on model selection methods via penalization. In both cases, the slope heuristic method of Chapter 6 is applied.

Each section of this document ends with a discussion and directions for future research. A complete list of our papers can be found at the end of the document, and all our papers can be downloaded here.

Contents

I Statistical aspects of topological data analysis **11**

1 Introduction **12**
1.1 Topological data analysis . . . 12
1.2 Approximating models for TDA: offsets and simplicial complexes . . . 13
1.3 Simplicial homology . . . 15
1.4 Topological inference and reconstruction procedures . . . 16
1.5 Statistical approaches to TDA and topological inference . . . 18

2 Model selection for simplicial approximation **20**
2.1 Geometric models . . . 20
2.2 Selection of a simplicial complex in the filtration . . . 21
2.3 Applications . . . 22
2.4 Discussion and directions for future research . . . 24

3 A statistical approach to persistent homology on metric spaces **25**
3.1 Persistence diagrams and persistence landscapes . . . 26
3.2 Estimation of persistent diagrams on metric spaces . . . 28
3.3 Subsampling methods for persistent homology . . . 30
3.3.1 The multiple samples approach . . . 31
3.3.2 Stability of the average landscape . . . 31
3.3.3 Risk analysis . . . 32
3.4 Experiments . . . 33
3.5 Discussion and directions for future research . . . 34

4 Robust topological data analysis with the distance to measure **35**
4.1 The distance to measure . . . 35
4.2 Rates of convergence of the DTEM . . . 37
4.2.1 Local analysis of the DTEM in the bounded case . . . 37
4.2.2 Local analysis of the DTEM in the unbounded case . . . 38
4.2.3 About the geometric information carried by the quantile function $\mathbf{F}_{\mathbf{x}}^{-1}$ . . . 39
4.3 Limiting distribution and bootstrap for the DTM . . . 40
4.3.1 Hadamard differentiability and bootstrap for the DTM . . . 40
4.3.2 Bootstrap and significance of topological features . . . 41
4.4 Denoising the DTM via Wasserstein deconvolution . . . 42
4.4.1 Deconvolution of a measure and Wasserstein metric . . . 43
4.4.2 Rates of convergence . . . 44
4.5 Discussion and directions for future research . . . 46

II Other contributions in the field of Statistics **48**

5 Gaussian mixture clustering **49**
5.1 Gaussian mixture selection through ℓ_0 penalization . . . 49
5.2 Minimax adaptivity in the unidimensional case . . . 51

5.3 Discussion and directions for future research 51

6 Slope Heuristics and the Capushe package **52**
6.1 Contrast minimization and slope heuristics 52
6.2 Dimension jump 54
6.3 Data-driven slope estimation method 55

7 Feature selection for Random Forests **57**
7.1 Random forests 57
7.2 Permutation importance measure and feature selection 57
7.3 Grouped variable importance measure 59
7.4 A case study: variable selection for aviation safety 59
7.5 Discussion and directions for future research 59

Publication list **61**

Bibliography **63**

Part I

Statistical aspects of topological data analysis

Chapter 1

Introduction

This chapter is an introduction to topological data analysis (TDA) and topological inference methods. The necessary background in topology, geometry and computational geometry is briefly recalled. A nonexhaustive presentation of classical results about topological reconstruction and topological inference is presented. The last section of the chapter gives the motivations behind a statistical approach to this subject, as developed in the following chapters.

In a given metric space $(\mathbb{M}, \rho)$, the closed ball centered at $x \in \mathbb{M}$ with radius r is denoted by $B(x, r)$. The Hausdorff distance between compact sets is denoted $\mathrm{d_H}$. In Euclidean spaces the metric is defined by the Euclidean norm $\|\cdot\|$. The transpose of a matrix A is denoted A^t. This notation is used in all this first part of the habilitation thesis.

1.1 Topological data analysis

During the previous decades, wide availability of measurement devices and simulation tools has led to an explosion in the amount of available data in almost all domains of science, industry, economy and even everyday life. Often these data come as point clouds sampled in possibly high (or infinite) dimensional spaces. They are usually not uniformly distributed in the embedding space but carry some geometric structure which reflects important properties of the "systems" from which they have been generated.

There exist various statistical and machine learning methods that aim to uncover the geometric structure of data, including clustering, manifold learning and nonlinear dimensionality reduction, principal curves and sets estimation, to name a few. Most of them assume the underlying structure to have a very simple geometry — homeomorphic to a disc or isometric to an open set of a Euclidean space. Furthermore the only topological information they look for is connectivity.

With the emergence of new geometric inference and algebraic topology tools, computational topology (Edelsbrunner and Harer, 2010) has recently witnessed important developments with regards to data analysis, giving birth to the field of topological data analysis (TDA), whose aim is to infer relevant, qualitative and quantitative topological structures directly from the data (Carlsson, 2009).

The field of topological data analysis actually refers to various approaches and methods for exploring data. The two most popular approaches in TDA are probably the Mapper algorithm (Singh et al., 2007) and persistent homology (Edelsbrunner et al., 2002). The Mapper algorithm is a visualization method that preserves topological structure, whereas persistent homology provides a framework and efficient algorithms to encode the evolution of the topology of shape from small to large scale. We do not study the Mapper algorithm in this habilitation thesis, but persistent homology is the main subject of Chapter 3, and to a lesser extent Chapter 4. One fundamental question underlying TDA is what kind of topological information can be extracted in practice from data. This problem corresponds to the field of topological inference.

Topological inference methods aim to infer topological properties of an unknown topological space. Typically, a point cloud $\mathbb{X}_n$ is observed and the data is supposed to have be sampled in a neighborhood of the unknown shape $\mathbb{X}$, as illustrated by Figure 1.1. For $\mathbb{X}_n$ "close" enough to $\mathbb{X}$, it is expected that the topology of $\mathbb{X}$ can be inferred from $\mathbb{X}_n$. In Euclidean spaces and more generally metric spaces,

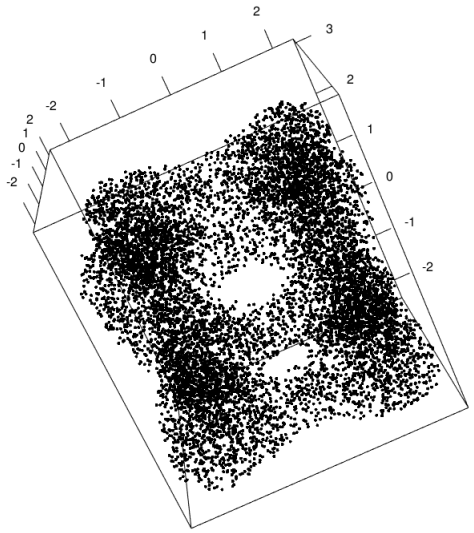


Figure 1.1: Point cloud sampled on the tangle cube in $\mathbb{R}^3$.

closeness between (compact) sets can described using various metrics. The traditional approach in topological inference requires closeness between compact sets for the Hausdorff distance between $\mathbb{X}$ and its approximation.

Generally speaking, point clouds in themselves do not carry any non-trivial topological or geometric structure. It is thus necessary to consider geometric structures on top of such point clouds in order to recover information about the shapes they approximate. In the most favorable situations, it is possible to define such an approximation $\tilde{\mathbb{X}}_n$ of the underlying object $\mathbb{X}$ that is homeomorphic to $\mathbb{X}$. In this case the topology of $\mathbb{X}$ can be described by the values of topological invariants on $\tilde{\mathbb{X}}_n$, as for instance the number of components, the homotopy type, its Betti numbers (see below), etc. A weaker relation between topological spaces which however preserves many topological invariants is the homotopy equivalence[1].

The next section introduces the approximating models used for topological inference and more generally for TDA.

1.2 Approximating models for TDA: offsets and simplicial complexes

One natural strategy to infer topological information for an unknown shape from a point cloud is to consider the *offsets* of the point cloud. For a point cloud $\mathbb{X}_n$ in $\mathbb{R}^d$ (or in a metric space), the r-offset of $\mathbb{X}_n$ is defined by

$$\mathbb{X}_n^r = \bigcup_{x \in \mathbb{X}_n} B(x, r).$$

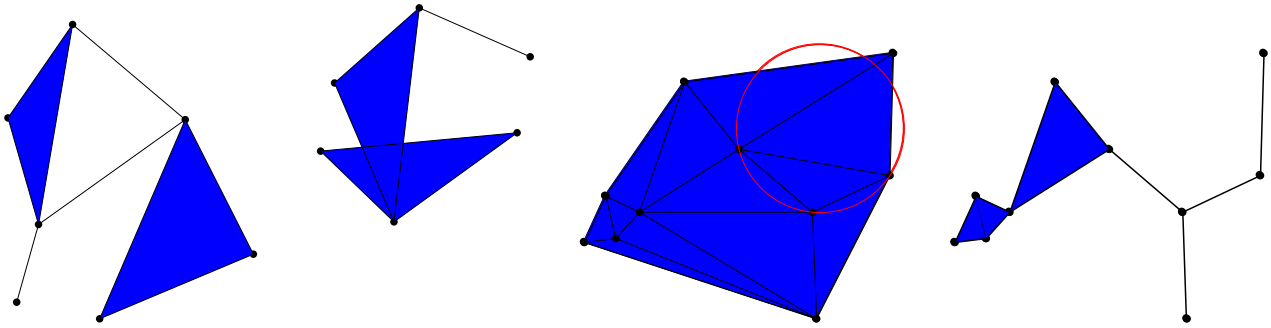
More generally, for any set $\mathbb{X}$ in a metric space $(\mathbb{M}, \rho)$, the r-offset $\mathbb{X}^r$ of $\mathbb{X}$ is defined by

$$\mathbb{X}^r = \bigcup_{x \in \mathbb{X}} B(x, r).$$

However, non-discrete sets such as offsets, and also continuous mathematical shapes like curves, surfaces and more generally manifolds, cannot easily be encoded as finite discrete structures. Simplicial complexes are therefore used in computational geometry to approximate such shapes. These can be seen as generalizations of neighborhood graphs.

The definition of geometric simplicial complexes in $\mathbb{R}^d$ is now recalled, see also Figure 1.2(a). A n-dimensional simplex s is the set of convex combinations of n+1 affinely independent points $\mathbb{X}_n =$

[1]Given two topological spaces $\mathbb{X}$ and $\mathbb{Y}$, two maps f_0, $f_1 : \mathbb{X} \to Y$ are homotopic if there exists a continuous map $H : [0,1] \times \mathbb{X} \to \mathbb{Y}$ such that for all $x \in \mathbb{X}$, $H(0,x) = f_0(x)$ and $H(1,x) = f_1(x)$. The two spaces $\mathbb{X}$ and $\mathbb{Y}$ are homotopy equivalent if there exist two continuous maps $f : \mathbb{X} \to \mathbb{Y}$ and $g : Y \to X$ such that $g \circ f$ is homotopic to the identity map in $\mathbb{X}$ and $f \circ g$ is homotopic to the identity map in $\mathbb{Y}$.



(a) The left set of simplices is a simplicial complexe whereas the right set of simplices is not.

(b) Delaunay complex in the plane and its empty ball characterization.

(c) one α-complex on the same set of points as for the Delaunay Complex.

Figure 1.2: Some geometric simplices in $\mathbb{R}^2$.

$\{x_0, \ldots, x_n\}$. The points x_i are called vertices. The simplices spanned by non-empty subsets of $\mathbb{X}_n$ are called faces of s. Note that a point is a 0-simplex, a segment is a 1-simplex and triangle is a 2-simplex.

Definition 1. *A geometric simplicial complex $\mathcal{C}$ is a set of simplices such that:*

- *Any face of a simplex from $\mathcal{C}$ is also in $\mathcal{C}$.*
- *The intersection of any two simplices s_1, $s_2 \in \mathcal{C}$ is either a face of both s_1 and s_2, or empty.*

A simplicial complex can also be seen as a combinatorial object consisting of subsets of the full vertex set of the complex. This remark motivates the following definition of abstract simplicial complexes.

Definition 2. *Let $\mathbb{X}_n = \{x_0, \ldots, x_n\}$ be a finite set of elements. An abstract simplicial complex $\mathcal{C}$ with vertex set $\mathbb{X}_n$ is a set of subsets of $\mathbb{X}_n$ such that:*

- *The elements of $\mathbb{X}_n$ belong to $\mathcal{C}$;*
- *If $s \in \mathcal{C}$ and $\varnothing \neq \tilde{s} \subset s$, then $\tilde{s} \in \mathcal{C}$.*

One important large family of constructions of simplicial complexes relies on the Delaunay complex. An exhaustive presentation of the Delaunay complex and its variants can be found for instance in Boissonnat et al. (2015). In this habilitation thesis, we only present the Delaunay complex and the α-shape complex, see also Figures 1.2(b) and 1.2(c). Let $\mathbb{X}_n = \{x_1, \ldots, x_n\}$ be a point cloud of $\mathbb{R}^d$ which is in general position[2].

- A simplex $[x_{i_0}, x_{i_1}, \cdots, x_{i_k}]$ is in the Delaunay complex of $\mathbb{X}_n$ if and only if it has a circumscribing ball empty of points of $\mathbb{X}_n$.
- A simplex $[x_{i_0}, x_{i_1}, \cdots, x_{i_k}]$ is in the α-complex of $\mathbb{X}_n$ if and only if $[x_{i_0}, x_{i_1}, \cdots, x_{i_k}]$ is in the Delaunay complex and the square radius of its circumscribing ball is at most α. The α-shape of $\mathbb{X}_n$ for scale parameter α is the set defined as the union of the simplices in the α-complex of $\mathbb{X}_n$.

The Delaunay complex and the α-complex are both embedded in $\mathbb{R}^d$ and they can be used for approximating an unknown shape in low dimension, typically for the Hausdorff metric. The α-complex can be also used for topological inference because it is homotopy equivalent to the union of the balls (Edelsbrunner, 1993). However, as computation of the Delaunay complex is limited for practical reasons to very low dimensions, alternative constructions need to be considered for topological inference, like for instance the Čech and Vietoris-Rips (or Rips) complexes. For a point cloud $\mathbb{X}_n = \{x_0, x_1, \cdots, x_n\}$ in $\mathbb{R}^d$ and $\alpha > 0$, let $\mathrm{Cech}_\alpha(\mathbb{X}_n)$ and $\mathrm{Rips}_\alpha(\mathbb{X}_n)$ be the Čech and Rips complexes of scale parameter α built on $\mathbb{X}_n$ (see also Figure 1.3):

- A simplex $[x_{i_0}, x_{i_1}, \cdots, x_{i_k}]$ is in the Čech complex $\mathrm{Cech}_\alpha(\mathbb{X}_n)$ if and only if $\bigcap_{j=0}^{k} B(x_{i_j}, \alpha) \neq \varnothing$.

[2] namely, no subset of $d+2$ points of $\mathbb{X}_n$ lies on the same hypersphere.

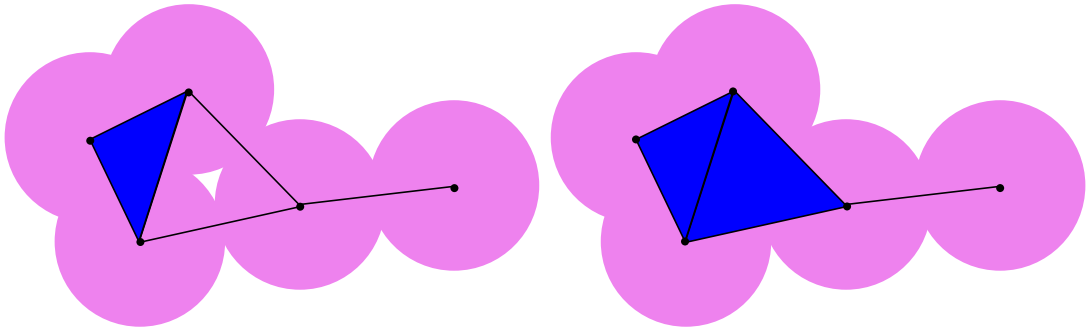


Figure 1.3: Left : Čech complex built on five points with a given scale parameter. Right: Rips complex built on the same points and with the same scale parameter α. The 2-simplices (or triangles) which lie in the complex are filled in blue. The offset $\mathbb{X}_n^\alpha$ is represented in pink.

- A simplex $[x_{i_0}, x_{i_1}, \cdots, x_{i_k}]$ is in the Rips complex $\text{Rips}_\alpha(\mathbb{X}_n)$ if and only if $\|x_{i_j} - x_{i_{j'}}\| \leqslant \alpha$ for all $j, j' \in \{1, \ldots, k\}$.

The definition of Čech and Vietoris-Rips complexes is not limited to the case of Euclidean spaces; they can be defined for a set of points in any metric space. In fact, the definition can be extended to any compact metric space (Chazal et al., 2014d).

The Nerve Theorem is a classical result in algebraic topology, see for instance Hatcher (2002). It says that the offsets $\mathbb{X}_n^\alpha$ of a point cloud $\mathbb{X}_n$ in $\mathbb{R}^d$ are homotopy equivalent to the Čech complex $\text{Cech}_\alpha(\mathbb{X}_n)$. This result opens the door to computational topology: the topology of the offsets can be inferred from the topology of Čech complexes. For instance, Betti numbers on simplicial complexes (defined in the next section) can be computed with efficient algorithms. However, computation of Čech complexes quickly becomes difficult when the dimension increases. In practice, Rips complexes can also be used for encoding the topology of the offsets because of the following property:

Proposition 1. *Let $\mathbb{X}_n$ be a set of points in $\mathbb{R}_n$. Then, for any $\alpha > 0$:*

$$\text{Rips}_\alpha(\mathbb{X}_n) \subset \text{Cech}_\alpha(\mathbb{X}_n) \subset \text{Rips}_{2\alpha}(\mathbb{X}_n).$$

Simplicial complexes are usually parametrized by a scale parameter and the complete collection of simplicial complexes is called a filtration. More formally:

Definition 3. *A filtration $(\mathcal{C}_k)_{k=0,\ldots,m}$ of a finite simplicial complex $\mathcal{C}$ is an increasing sequence of sub-complexes such that*

- $\varnothing = \mathcal{C}_0 \subset \mathcal{C}_1 \subset \cdots \subset \mathcal{C}_m = \mathcal{C}$,
- $\mathcal{C}_{k+1} = \mathcal{C}_k \cup s^{k+1}$ *where* s^{k+1} *is a simplex of* $\mathcal{C}_{k+1}$.

For instance, the family of α-complexes is a filtration of the Delaunay complex. When simplicial complexes depend on a scale parameter α, by abuse of definition the filtration can also be indexed by the scale parameter α: the filtration $(\mathcal{C}_\alpha)_{\alpha \in [0,\bar{\alpha}]}$ is a filtration of the complex $\mathcal{C}_{\bar{\alpha}}$.

1.3 Simplicial homology

Homology is a mathematical formalism used to summarize connected components, holes, tunnels and voids in general in a topological space. The definition of homology on simplicial complexes, that is *simplicial homology*, is now briefly recalled. A complete presentation of simplicial homology and singular homology can be found for instance in Munkres (1984) or in Hatcher (2001).

In this habilitation thesis, we only consider simplicial homology on $\mathbb{Z}/2\mathbb{Z}$. In this framework, simplicial homology has an obvious topological and geometric interpretation. Let $\mathcal{C} = \{s_1, \ldots, s_k\}$ be a simplicial complex and k a dimension. A k-chain c is a formal sum of k-simplicial complexes in $\mathcal{C}$: $c = \sum a_i s_i$, where coefficients a_i are taken from $\mathbb{Z}/2\mathbb{Z}$, and where the chain can be though of as

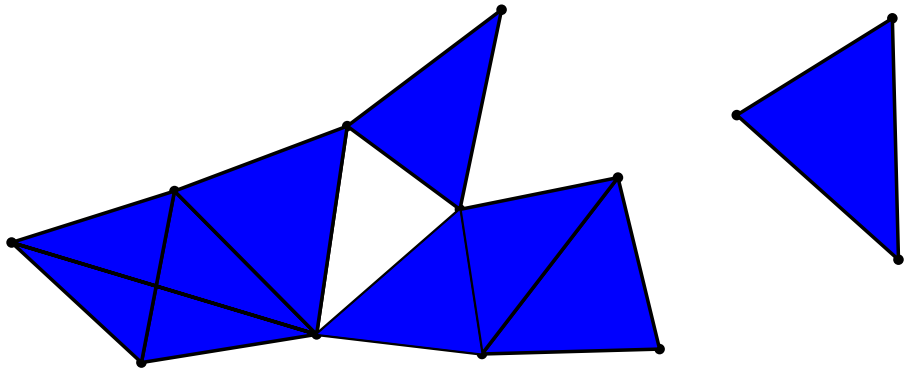


Figure 1.4: Simplicial homology example: for this simplicial complex, $\beta_0 = 2$ and $\beta_1 = 1$.

composed of the simplices whose coefficients are equal to one. A sum of two k chains $c = \sum a_i s_i$ and $c' = \sum b_i s_i$ is defined by $c + c' = \sum (a_i + b_i) s_i$, where the sum between a_i and b_i is the sum in $\mathbb{Z}/2\mathbb{Z}$. This can be seen as the symmetric difference of the two chains. With this addition operation, the set of k chains on $\mathcal{C}$, denoted by $C_k(\mathcal{C})$, is an abelian group.

The boundary $\partial_k s$ of a k-simplex s is defined as the sum of its $k-1$ faces and the boundary $\partial_k c$ of a k-chain $c = \sum a_i s_i$ is defined as the sum of the boundaries of its simplexes. The elements of the subgroup $Z_p := \mathrm{Ker}\partial_k$ are called *cycles*: a k-cycle c is a k-chain with empty boundary. The elements of the subgroup $B_p := \mathrm{Im}\partial_{k+1}$ are the k-boundaries: a k-boundary c is the boundary of a $(k+1)$-chain. The main property of this construction is that the boundary of a boundary is necessarily zero. Intuitively, the homology groups of $\mathcal{C}$ correspond to the voids of dimension k of the simplicial complex (see Figure 1.4 for an illustration).

Definition 4. *The k-th homology group of $\mathcal{C}$ is the k-th cycle group modulo the k-th boundary group: $H_p = Z_p/B_p$. The k-th Betti number is the rank of this group: $\beta_k = \mathrm{rank}\ H_p$.*

In practice, one build a simplicial complex on a point cloud and Betti numbers appear as simple and interpretable topological signatures of the underlying shape on which the point cloud has been sampled. The notion of homology can be extended to general topological spaces by considering *singular homology*. This notion is beyond the scope of the thesis. We only mention here the key fact that singular homology is a topological invariant. This last property is, in some sense, the justification for using homology and Betti numbers computed on simplicial complexes for describing an unknown shape.

1.4 Topological inference and reconstruction procedures

This section gives a short presentation of topological inference results, see Boissonnat et al. (2015) for more details. There are two main facts relevant to topological inference results. The first is that, unsurprisingly, the difficulty in inferring the topology of a shape directly depends on its "regularity". There are several ways to quantify the regularity of a geometric shape. The second is that a complete theory of topological inference can be derived from the study of distance functions to compact sets.

For a compact set $\mathbb{X}$ in $\mathbb{R}^d$, the distance function $d_{\mathbb{X}}$ to $\mathbb{X}$ is the non-negative function defined by

$$d_{\mathbb{X}}(y) = \inf_{x \in \mathbb{X}} \|x - y\|.$$

Note that $\mathbb{X}$ is completely characterized by $d_{\mathbb{X}}$ since $\mathbb{X} = d_{\mathbb{X}}^{-1}(0)$. Moreover the r-offset $\mathbb{X}^r$ can be defined by $\mathbb{X}^r = d_{\mathbb{X}}^{-1}([0, r])$. For some point y in the complement $\mathbb{X}^c$ of $\mathbb{X}$, let $\Gamma(y)$ be the set of points in $\mathbb{X}$ closest to y: $\Gamma_{\mathbb{X}}(y) = \{x \in \mathbb{X} \mid \|x - y\| = d_{\mathbb{X}}(y)\}$. The medial axis of $\mathbb{X}^c$ is defined by $\mathcal{M}(\mathbb{X}^c) = \{y \in \mathbb{X}^c, |\Gamma_{\mathbb{X}}(y)| \geqslant 2\}$. Several regularity properties of geometric shapes can be expressed as a function of the medial axis of its complement.

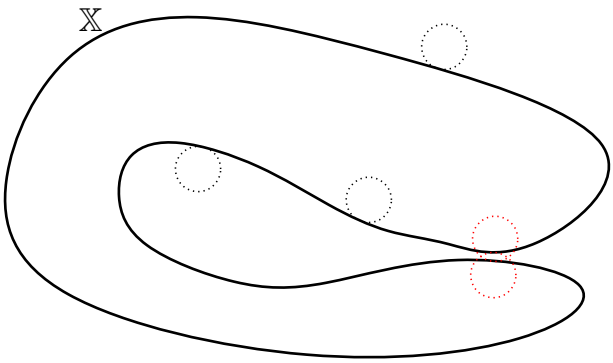


Figure 1.5: This manifold is close to being self-intersecting and has a small reach.

Local feature size and reach

A local notion of regularity of a compact set $\mathbb{X}$ can be measured by the so-called *local feature size*. For $x \in \mathbb{X}$, it is defined by $\mathrm{lfs}_{\mathbb{X}}(x) := d\left(x, \mathcal{M}(\mathbb{X}^c)\right)$. The global version of the local feature size is the *reach*[3] introduced by Federer (1959): $\kappa(\mathbb{X}) = \inf_{x \in \mathbb{X}^c} \mathrm{lfs}_{\mathbb{X}}(x)$. The reach is small if either $\mathbb{X}$ is not smooth or if $\mathbb{X}$ is close to being self-intersecting (see Figure 1.5) and the same remark is of course also true locally for the local feature size. Amenta et al. (2000) show that a topologically correct reconstruction of a surface smoothly embedded in $\mathbb{R}^3$ is possible from a point cloud, as soon as every point $x \in \mathbb{X}$ has a sample point at distance at most $0.06\,\mathrm{lfs}_{\mathbb{X}}(x)$. However, this result cannot be applied when the geometric shape has sharp edges because the local feature size vanishes on such edges.

Weak feature size and its extensions

The *weak feature size* is a more flexible notion of regularity then the reach. It relies on the notion of critical points for $d_{\mathbb{X}}$. The function $d_{\mathbb{X}}$ is not differentiable everywhere but a generalized gradient vector field $\nabla d_{\mathbb{X}}$ for $d_{\mathbb{X}}$ can be defined as follows:

$$\nabla d_{\mathbb{X}}(x) = \begin{cases} \frac{x-\theta}{d_{\mathbb{X}}(x)} & \text{if } x \notin \mathbb{X} \\ 0 & \text{if } x \in \mathbb{X}, \end{cases}$$

where θ is the center of the smallest closed ball enclosing $\Gamma(x)$, see Figure 1.6.

Definition 5. *A point x is a critical point of $d_{\mathbb{X}}$ if $\nabla d_{\mathbb{X}}(x) = 0$. A real $c \geqslant 0$ is a critical value of $d_{\mathbb{X}}$ if there exists a critical point $x \in \mathbb{R}^d$ such that $d_{\mathbb{X}}(x) = c$. A regular value of $d_{\mathbb{X}}$ is a value which is not critical.*

The weak feature size of a geometric shape was introduced in Chazal and Lieutier (2007):

Definition 6. *The weak feature size* $\mathrm{wfs}(\mathbb{X})$ *of $\mathbb{X}$ is the infimum of the positive critical values of $d_{\mathbb{X}}$. If $d_{\mathbb{X}}$ does not have critical values then* $\mathrm{wfs}(\mathbb{X}) = +\infty$.

Using the notion of critical point, Grove (1993) has shown that the sublevel sets of $d_{\mathbb{X}}$ are topological submanifolds of $\mathbb{R}^d$ and that their topology can change only at critical points. Moreover, for $0 < \alpha < \beta < \mathrm{wfs}(\mathbb{X})$, the offsets $\mathbb{X}^\alpha$ and $\mathbb{X}^\beta$ are isotopic[4].

The following theorem is a typical example in the literature about topological inference of *stability result*. It shows that for some range of values of the scale parameter, two close compact sets have the same offset topology.

Theorem 1. [Chazal and Lieutier 2007] *Let $\mathbb{X}$ and $\mathbb{Y}$ be two compact sets in $\mathbb{R}^d$ and let $\varepsilon > 0$ be such that $\mathrm{d}_H(\mathbb{X}, \mathbb{Y}) < \varepsilon$, $\mathrm{wfs}(\mathbb{X}) > 2\varepsilon$ and $\mathrm{wfs}(\mathbb{Y}) > 2\varepsilon$. Then for any $0 < \alpha < 2\varepsilon$, $\mathbb{X}^\alpha$ and $\mathbb{Y}^\beta$ are homotopy equivalent.*

[3] also called *condition number*

[4] The notion of isotopy is stronger than homeomorphy to distinguish between spaces in $\mathbb{R}^d$. An isotopy between $\mathbb{X}$ and $\mathbb{Y}$ is a continuous application $F\colon \mathbb{X} \times [0,1] \to \mathbb{R}^d$ such that $F(.,0)$ is the identity map on $\mathbb{X}$, $F(\mathbb{X},1) = \mathbb{Y}$ and for any $t \in [0,1]$, F is an homeomorphism between $\mathbb{X}$ and $F(\mathbb{X},t)$.

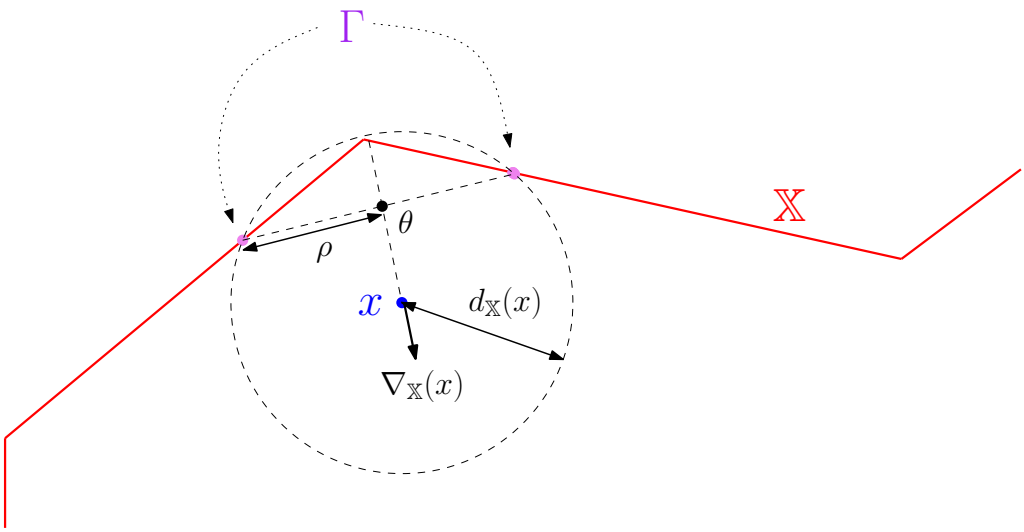


Figure 1.6: Definition a generalized gradient for the distance function to a compact set.

However, the assumptions of Theorem 1 are not satisfied in the realistic case where an unknown shape $\mathbb{X}$ is approximated by a point cloud $\mathbb{X}_n$. Indeed, the weak feature size of a finite point cloud $\mathbb{X}_n$ is equal to half of the distance between the two closest points of $\mathbb{X}_n$. In most cases, the two conditions $\mathrm{d_H}(\mathbb{X}, \mathbb{X}_n) < \varepsilon$ and $\mathrm{wfs}(\mathbb{X}_n) > 2\varepsilon$ are not simultaneously satisfied.

To deal with this drawback, improvements to Theorem 1 have been proposed in particular in Chazal et al. (2009c). In short, a more general notion of regularity is introduced: the *μ-reach*, which interpolates between the minimum of the local feature (the reach) and the weak feature size. By considering this quantity as a measure of the regularity of the shape, it is shown in Chazal et al. (2009c) that the homotopy type[5] of $\mathbb{X}$ or at least that of its small offsets can be inferred from the homotopy types of the offsets of an approximation of $\mathbb{X}$. Here, the proximity required between $\mathbb{X}$ and its approximation is in terms of the Hausdorff distance and also depends on the μ-reach of $\mathbb{X}$.

A first probabilistic statement of topological reconstruction

In the paper Niyogi et al. (2008), it is shown that the homotopy type of Riemannian manifolds with reach larger than a given constant can be recovered with high probability from offsets of a sample on (or close to) the manifold. This paper is very important in the computational geometry literature since it was the first paper to consider the topological inference problem in terms of probability. The result of Niyogi et al. (2008) is derived from a retract contraction argument and on tight bounds over the packing number of the manifold in order to control the Hausdorff distance between the manifold and the observed point cloud. The assumption that the geometric object is a smooth Riemannian manifold is only used in the paper to control in probability the Hausdorff distance between the sample and the manifold, and not actually necessary for the "topological part" of the result. Regarding the topological results, these are similar to those of Chazal et al. (2009c) in the particular framework of Riemannian manifolds. Starting from the result of Niyogi et al. (2008), the minimax rates of convergence of the homology type have been studied by Balakrishnan et al. (2012) under various models, for Riemannian manifolds with reach larger than a constant. In contrast, a statistical version of Chazal et al. (2009c) has not yet been proposed.

1.5 Statistical approaches to TDA and topological inference

Until very recently, the theory on TDA and topological inference mostly relied on deterministic approaches, as presented above. These deterministic approaches do not take into account the random nature of data and the intrinsic variability of the topological quantity they infer. Consequently, most

[5] actually it is also true for the isotopy type, see Chazal et al. (2009b).

of the corresponding methods remain exploratory, without being able to efficiently distinguish between information and what is sometimes called the "topological noise".

A statistical approach to TDA means that we consider data as generated from an unknown distribution, but also that the inferred topological features by TDA methods are seen as estimators of topological quantities describing an underlying object. Under this approach, the unknown object corresponds to the support of the data distribution (or at least is close to this support). However, this support does not always have a physical existence; for instance, galaxies in the universe are organized along filaments but these filaments do not physically exist. A statistical approach to TDA is thus strongly related to the problem of *distribution support estimation* and *level sets estimation*[6] under the Hausdorff metric, as suggested by the stability results presented in the previous section.

A large number of methods and results are available for estimating the support of a distribution in statistics. For instance, the Devroye and Wise estimator (Devroye and Wise, 1980) defined on a sample $\mathbb{X}_n$ is also a particular offset of $\mathbb{X}_n$. The convergence rates of both $\mathbb{X}_n$ and the Devroye and Wise estimator to the support of the distribution for the Hausdorff distance is studied in Cuevas and Rodríguez-Casal (2004) in $\mathbb{R}^d$. More recently, the minimax rates of convergence of manifold estimation for the Hausdorff metric, which is particularly relevant for topological inference, has been studied in Genovese et al. (2012). There is also a large literature about level sets estimation in various metrics (see for instance Polonik, 1995; Tsybakov et al., 1997; Cadre, 2006) and more particularly for the Hausdorff metric in Chen et al. (2015). All these works about support and level sets estimation shine light on the statistical analysis of topological inference procedures.

The main goals of a statistical approach to topological data analysis can be summarized as the following list of problems:

Topic 1: proving consistency and studying the convergence rates of TDA methods.

Topic 2: providing confidence regions for topological features and discussing the significance of the estimated topological quantities.

Topic 3: selecting relevant scales at which the topological phenomenon should be considered, as a function of observed data.

Topic 4: dealing with outliers and providing robust methods for TDA.

The following chapters in this part of the thesis present our contributions to this statistical approach to TDA. The immediately following chapter gives a model selection method for automatically selecting a simplicial complex in a given filtration; this corresponds to Topic 3. Chapter 3 is about statistical methods for the estimation of persistence diagrams; the contributions of this chapter provide some answers to Topics 1 and 2. Chapter 4 is on the statistical analysis of a robust method for TDA based on the distance to measure. The contributions of this chapter correspond to Topics 1, 2 and 4.

[6] In particular for persistent homology, see for instance Fasy et al. (2014)

Chapter 2

Model selection for simplicial approximation

Given a point cloud $\mathbb{X}_n$ and a filtration of simplicial complexes $(\mathcal{C}_\alpha)_{\alpha\in\mathcal{A}}$, choosing a convenient scale parameter for topological inference or for reconstruction is not obvious. In this chapter, we address the problem of selecting a "convenient" simplicial complex as a model selection problem, as proposed in our paper Caillerie and Michel (2011). Our method relies on the theory of non-asymptotic model selection by penalization.

In this chapter, for $q \in \mathbb{N}^*$, the space $\mathbb{R}^q$ is equipped with the following normalized scalar product :

$$\forall u, v \in \mathbb{R}^q, \quad \langle u, v\rangle_{[q]} := \frac{1}{q}\sum_{i=1}^{q} u_i v_i \ , \tag{2.1}$$

and the associated norm is denoted $\|\cdot\|_{[q]}$.

2.1 Geometric models

In the standard setting of topological inference in $\mathbb{R}^d$, an unknown geometric object $\mathbb{X}$ embedded in $\mathbb{R}^d$ is approximated from a point cloud $\mathbb{X}_n$ which points are observed in the neighborhood of $\mathbb{X}$. We then assume that the observed points $X_1, \dots, X_n$ satisfy

$$\forall i = 1, \dots, n, \quad X_i = \bar{x}_i + \sigma\xi_i \quad \text{with} \quad \bar{x}_i \in \mathbb{X}, \tag{2.2}$$

where the original points $\bar{x}_i$ are unknown and the random variables ξ_i are independent standard Gaussian vectors of $\mathbb{R}^d$ and σ is the noise level. Let $\mathbf{X} = (X_1^t, \dots, X_n^t)^t$ be the vector of length $q = nd$ containing all the observations X_i of the point cloud $\mathbb{X}_n$. We also define $\bar{\mathbf{x}}$ and $\boldsymbol{\xi}$ in the same way. We consider the next equivalent statement of (2.2) in the space $\mathbb{R}^{nd}$:

$$\mathbf{X} = \bar{\mathbf{x}} + \sigma\boldsymbol{\xi} \quad \text{with} \quad \bar{\mathbf{x}} \in \mathbb{X}^{nd}, \tag{2.3}$$

where $\boldsymbol{\xi}$ is a standard Gaussian vector of $\mathbb{R}^{nd}$.

In this work, we consider the geometric realization of a simplicial complex: by simplicial complexes we actually mean the support of the complexes by abuse of definition. For a given simplicial complex $\mathcal{C}$, the best approximating point of $\bar{\mathbf{x}}$ belonging to $\mathcal{C}$ minimizes the quantity $\mathbf{t} \mapsto \|\mathbf{t} - \bar{\mathbf{x}}\|_{[nd]}$. The least square estimator (LSE) of $\bar{\mathbf{x}}$ associated to the complex $\mathcal{C}$ is then defined by

$$\hat{\mathbf{x}} := \operatorname{argmin}_{\mathbf{t}\in\mathcal{C}^n} \|\mathbf{X} - \mathbf{t}\|^2_{[nd]}, \tag{2.4}$$

where $\mathcal{C}^n$ denotes the Cartesian product of $\mathcal{C}$. For each $i = 1, \dots, n$, let $\hat{x}_i$ be the closest point of X_i belonging to $\mathcal{C}$; then $\hat{\mathbf{x}} = (\hat{x}_1^t, \dots, \hat{x}_n^t)^t$.

2.2 Selection of a simplicial complex in the filtration

Roughly speaking, a basic complex with only a few simplices will badly approximate $\mathbb{X}$ and the same is true for $\bar{\mathbf{x}}$, whereas a complex composed of too many simplices will tend to overfit the data. This fact corresponds in statistics to the well known *bias-variance trade off* and it can be figured out by model selection methods.

Let $\mathbb{P}_{\bar{\mathbf{x}}}$ be the distribution of $\mathbb{X}$ in (2.3) and let $(\mathcal{C}_\alpha)_{\alpha\in\mathcal{A}}$ be a filtration of simplicial complexes. We denote by $(C_\alpha := \mathcal{C}_\alpha^n)_{\alpha\in\mathcal{A}}$ the countable collection of Cartesian products of simplicial complexes and by $\hat{\mathbf{x}}_\alpha$ the LSE estimator corresponding to $\mathcal{C}_\alpha$, as defined in (2.4). The l^2-risk of $\hat{\mathbf{x}}_\alpha$ is defined by

$$\mathcal{R}(\bar{\mathbf{x}},\alpha) = \mathbb{E}_{\bar{\mathbf{x}}}\left(\|\bar{\mathbf{x}} - \hat{\mathbf{x}}_\alpha\|_{[nd]}^2\right),$$

where $\mathbb{E}_{\bar{\mathbf{x}}}$ is the expectation relative to $\mathbb{P}_{\bar{\mathbf{x}}}$. Ideally, we would like to choose the model $\alpha(\bar{\mathbf{x}})$ minimizing the risk: $\alpha(\bar{\mathbf{x}}) = \operatorname{argmin}_{\alpha\in\mathcal{A}}\mathcal{R}(\bar{\mathbf{x}},\alpha)$. The model $\alpha(\bar{\mathbf{x}})$ and the quantity $\hat{\mathbf{x}}_{\alpha(\bar{\mathbf{x}})}$, which is called *oracle*, are both unknown in practice but it is considered as a benchmark for theory. One popular method to select an estimator in a given family is *penalization*. In our context, this procedure consists of considering some proper penalty function $\text{pen} : \alpha \in \mathcal{A} \mapsto \text{pen}(\alpha) \in \mathbb{R}^+$ and of selecting $\hat{\alpha}$ minimizing the associated l^2 penalized criterion

$$\text{crit}(\alpha) = \|\mathbf{X} - \hat{\mathbf{x}}_\alpha\|_{[nd]}^2 + \text{pen}(\alpha). \tag{2.5}$$

The resulting selected estimator is denoted $\hat{\mathbf{x}}_{\hat{\alpha}}$. Obviously, the main difficulty of this approach is to choose a convenient penalty in order to select an estimator close to the oracle. For instance, the well known AIC penalty is $2d_\alpha\hat{\sigma}^2/n$ where $\hat{\sigma}^2$ is an estimator of the noise variance and d_α the "number of parameters" estimated by $\hat{\mathbf{x}}_{\hat{\alpha}}$. There is no obvious "number of parameters" associated to an estimator $\hat{\mathbf{x}}_{\mathcal{C}}$. The classical methods of penalization cannot be easily applied in our context.

A exhaustive theory of penalization with a non-asymptotic approach has been developed in the nineties, with the works of Birgé and Massart among others. This approach to model selection provides a penalty function leading to a oracle inequality for the penalized estimator. In Birgé and Massart (2001), such a non-asymptotic model selection result is obtained for collections of linear Gaussian models, namely if the C_α's were linear subspaces. For the case of nonlinear Gaussian models, Massart (2007) shows that efficient penalties can still be defined by using the metric entropy (Section 4.4 in in this book). We follow this approach for selecting simplicial complexes.

For a k-simplex s in $\mathbb{R}^d$, let Δ_s be the diameter of the smallest enclosing ball of s for the normalized norm (2.1) in $\mathbb{R}^d$. A simplicial complex is said to be k-homogeneous if each one of its simplices is either a k-simplex, or the face of a k-simplex of $\mathcal{C}$. Then, for a k-homogeneous simplicial complex $\mathcal{C}$ in $\mathbb{R}^d$, let $|\mathcal{C}|_k := (\sum_{s\in\mathcal{C}^+}\Delta_s^k)^{1/k}$ and $\delta_{\mathcal{C}} := \inf_{s\in\mathcal{C}^+}\Delta_s$ where $\mathcal{C}^+$ is the subset of simplices of $\mathcal{C}$ of maximal dimension k.

Let $\mathbf{X}$ be the observation vector with the distribution defined by (2.3). Let $(\mathcal{C}_\alpha)_{\alpha\in\mathcal{A}}$ be a given collection of k-homogeneous simplicial complexes in $\mathbb{R}^d$ and for each $\alpha \in \mathcal{A}$ let $\hat{\mathbf{x}}_\alpha$ be the LSE corresponding to $C_\alpha = \mathcal{C}_\alpha^n$. Assume that there exist some weights w_α such that $\sum_{\alpha\in\mathcal{A}} e^{-w_\alpha} = \Sigma < \infty$.

Theorem 2. [Caillerie and Michel 2011] *Under the previous hypotheses, also suppose that for all $\alpha \in \mathcal{A}$,*

$$\sigma \leqslant \delta_{\mathcal{C}_\alpha}\sqrt{\frac{d}{k}}\left[4\kappa\left(\sqrt{\ln\frac{4|\mathcal{C}_\alpha|_k}{\delta_{\mathcal{C}_\alpha}}} + \sqrt{\pi}\right)\right]^{-1}. \tag{2.6}$$

There exist some absolute constants c_1 and c_2 such that for all $\eta > 1$, if

$$\text{pen}(\alpha) \geqslant \eta\,\sigma^2\left(c_1\frac{k}{d}\left[\ln\frac{|\mathcal{C}_\alpha|_k\sqrt{d}}{\sigma\sqrt{k}} + c_2\right] + 4\frac{w_\alpha}{nd}\right), \tag{2.7}$$

then, almost surely, there exists a minimizer $\hat{\alpha}$ of the penalized criterion (2.5) and the penalized estimator $\hat{\mathbf{x}}_{\hat{\alpha}}$ satisfies the following risk bound

$$\mathbb{E}_{\bar{\mathbf{x}}}\|\hat{\mathbf{x}}_{\hat{\alpha}} - \bar{\mathbf{x}}\|_{[nd]}^2 \leqslant c_\eta\left[\inf_{\alpha\in\mathcal{A}}\left\{d(\bar{\mathbf{x}},\mathcal{C}_\alpha^n)^2 + \text{pen}(\alpha)\right\} + \frac{\sigma^2}{nd}(\Sigma+1)\right] \tag{2.8}$$

where c_η depends only on η and $d(\bar{\mathbf{x}},C_\alpha^n) := \inf_{\mathbf{y}\in C_\alpha^n}\|\bar{\mathbf{x}} - \mathbf{y}\|_{[nd]}$.

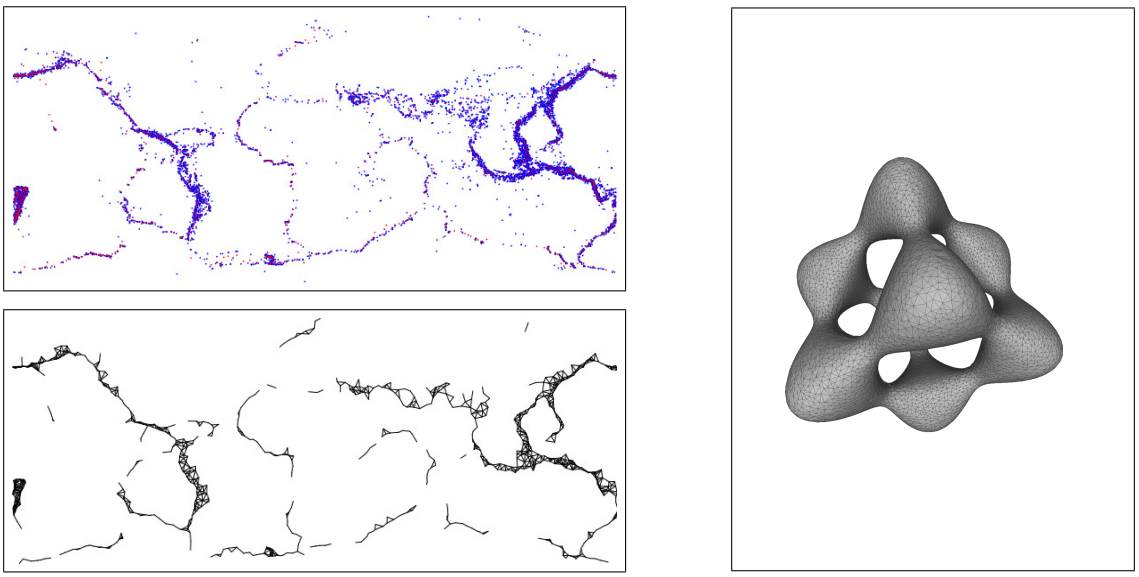


Figure 2.1: Landmarks in red and observed points in blue (top left) and selected α-complex (bottom left) for the seismic data. Selected α-complex for the Tangle Cube from noisy observations (right).

Condition (2.6) means that the complexes in the collection should not contain any k-simplexes with a diameter of the order of the noise level σ. This is natural since it would not be relevant to fit some simplices of this small scale on the observed data. This also means that the landmarks used to define the complexes should not be chosen too close to each other. The constant κ in the upper bound is an absolute constant which comes from Theorem 3.18 in Massart (2007).

The shape of the penalty function given by (2.7) is quite different than penalty shapes used in previous model selection works in the spirit of the results initiated by Birgé and Massart. The relevant term in the penalty bound (2.7) is the "size measurement" $\ln |\mathcal{C}_\alpha|_k$ of the complex. The penalty also depends on the weights w_α. By analogy with the case of linear models (see Massart, 2007, p.91), we can choose weights such that $w_\alpha = L \ln |\mathcal{C}_\alpha|_k$ with $\sum_{\alpha \in \mathcal{A}} \frac{1}{|\mathcal{C}_\alpha|_k^L} = \Sigma < \infty$, where $L > 0$. The lower bound (2.7) is then proportional to $\ln |\mathcal{C}_\alpha|_k$.

Note that bounds have no interest for the practice since they are surely far from being optimal. This theorem has to be considered from a qualitative point of view: the main contribution here is giving the penalty shape. This penalty shape does not directly depend on the geometric and topological properties of the complexes, but it is actually natural since the penalty is defined via the metric complexity of the simplicial complexes. However the method provides a "convenient scale" at which the geometric features have to be studied.

2.3 Applications

In practice, we consider α-complexes filtrations and we use the slope heuristics method presented in Chapter 6 to calibrate the penalty given in Theorem 2. In the particular case of graphs ($k = 1$), the term $\ln |\mathcal{C}_\alpha|_k$ exactly corresponds to the logarithm of the graph length, which is easy to compute. Various applications are proposed in Caillerie and Michel (2011). Figure 2.1 illustrates two applications of the method: one on a seismic dataset and one for the reconstruction on the Tangle Cube.

We also apply the method for spectral clustering, a popular clustering method based on the spectral decomposition of a matrix associated to a similarity graph (see for instance von Luxburg, 2007). The algorithm requires the choice of a similarity function, and a type of graph to define a similarity graph. A reasonable candidate for the similarity function is $s(x, y) = \exp(-\|x - y\|^2/(2\sigma^2))$. As to the graph, the k-nearest neighbor graph and the ε-neighborhood graph are mostly used in practice. However, as explained in von Luxburg (2007), choosing ε or k is a difficult question. As far as we known, there is no completely data-driven method to do this choice and no theoretical results is available to help the user. Furthermore, this choice has a deep impact on the clustering, as illustrated by the example

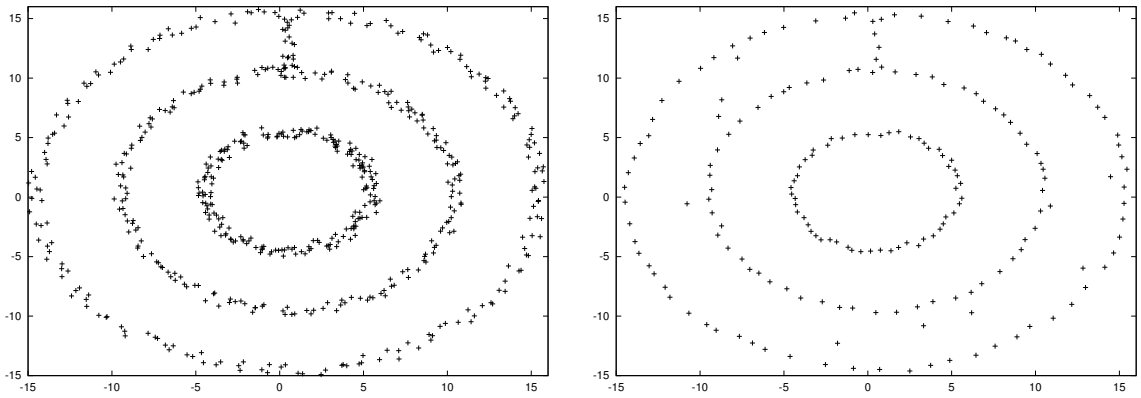


Figure 2.2: Data set (left) and 200 landmarks points (right).

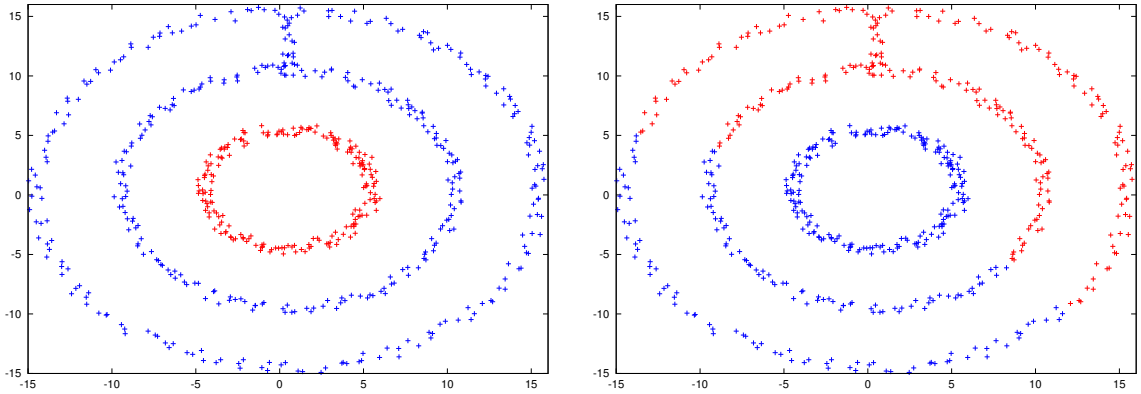


Figure 2.3: Classical spectral clustering based on a k-nearest neighbor with $k = 25$ (left) and $k = 30$ (right) : the clustering depends on k.

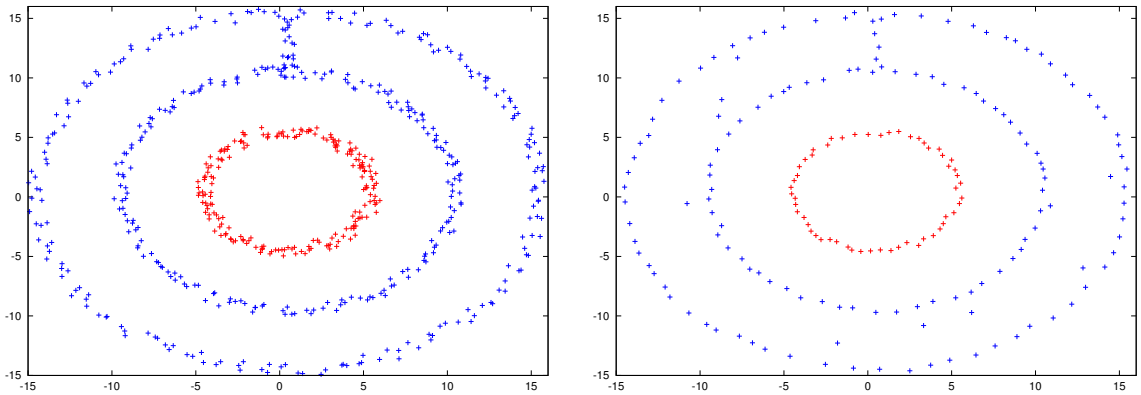


Figure 2.4: Spectral clustering based on the graph selection for the initial data points (left) and the landmark points (right). The labels exactly corresponds to the expected clustering.

below (see Fig. 2.2 and 2.3). To answer this problem, we select a graph on the data according to our method and next we proceed the spectral clustering method with the selected graph, see Fig. 2.4.

2.4 Discussion and directions for future research

Selection of heterogeneous simplicial complexes.

This model selection method does not prevent us to deal with heterogeneous simplicial complexes. The problem is that no explicit shape for the penalty can be proposed in the heterogeneous case because the equation involved in the penalty definition is much more complicated than in the homogeneous case. This difficult and interesting question should be tackled in future works.

About the landmarks.

In practice, the hypotheses of Theorem 2 are not completely satisfied since the computed complexes necessary depend on the observed data and thus are "not fixed" as in the theorem statement. Our theoretical result can be considered as conditional to the landmark choice. Giving some mathematical results for the "random models" we use would be much more difficult among other things because the distribution of the landmarks cannot be easily specified.

Data driven topological inference.

Our method is a completely data driven model selection method for approximating a shape for the ℓ_2 norm. By contrast, it does not directly answer to the problem of selecting a convenient scale in a filtration for topological inference. One first direction of research would be to study the performance of our method on topological inference problems. It was noticed in Chapter 1 that topological inference can be derived from proximity for the sup norm metric. We intend to adapt Lepski's methods for selecting a convenient scale in the filtration for the sup-nom metric.

Regarding the problem of estimating the homology of an underlying object, it must be noted that it is still not known how to build a reconstruction having the correct homology groups, in a data driven way. Consistency results have been proved by Niyogi et al. (2008) and more recently by Bobrowski et al. (2014), but these results are asymptotical. Consequently, the tuning of the scale parameters (or the bandwidth of the kernel estimators) proposed in these methods have no reason to be optimal in a non asymptotical point of view. Moreover they depend on geometric quantities which are unknown in practice. Finally, it is still unknown how to choose efficiently the scale parameters for a given point cloud of finite size. We would like to revisit the works of Niyogi et al. (2008) and Bobrowski et al. (2014) with model selection approaches in order to obtain a more data driven estimation method of the topology. The reach being a keystone quantity for all these methods, we currently study the estimation of this quantity from a statistical point, in a joint work in progress with F. Chazal, J. Kim, L. Wasserman and A. Rinaldo.

An alternative line of research about data driven topological inference would be to revisit the reconstructions results of Chazal et al. (2009c) with a statistical point of view. In this paper, a *critical function* is introduced which describes the regularity of the distance function $d_{\mathbb{X}}$ and some insights are also proposed on how to select the scale parameter in function of the critical function of the sample. Some stability results are also proven for the critical function. It would be interesting to study the convergence of the "empirical critical function", that is the critical function for $\mathbb{X}_n$ and to provide confidence regions for this last. This would open the door to a more "data driven" topological inference method, with statistical guarantees.

In the next chapter, we study the statistical aspects of *persistent homology*, an alternative approach to topological inference which consists in considering the complete filtration of simplicial complexes instead of considering only one particular scale.

Chapter 3

A statistical approach to persistent homology on metric spaces

In this chapter, we study the statistical aspects of *persistent homology*, one of the most popular approach in topological data analysis. We saw in the previous chapters that inferring the exact topology of an unknown shape, or at least of its small of sets, require some geometric regularity assumptions on the shape which can be hardly checked in practice. We also noticed in the discussion section at the end of the previous chapter that, even if the shape is smooth, selecting a convenient scale (in the filtration of simplicial complexes) for inferring the homology from a given point cloud, is a tricky problem. On the contrary, persistent homology provides multiscale topological information and it is not restricted to particular smooth geometric objects ; it can be actually used for any compact metric space.

Generally speaking, persistent homology comes with a theory (Edelsbrunner et al., 2002; Zomorodian and Carlsson, 2005; Edelsbrunner and Harer, 2010; Chazal et al., 2012) and efficient algorithms to encode the evolution of the homology of families of nested topological spaces indexed by a set of real numbers. In most cases it is computed for a filtered simplicial complex built on top of the available data, see Fig. 3.1 for an illustration. The obtained multiscale topological information is then represented in a simple way as a barcode, a *persistence diagram* (see Fig. 3.2) or a *persistence landscape* (see Fig. 3.3). These "topological signatures" are then used to exhibit and compare the topological structure underlying the data. Persistent homology has found applications in many fields, including neuroscience (Singh et al., 2008), bioinformatics (Kasson et al., 2007), shape classification (Chazal et al., 2009a), clustering (Chazal et al., 2013) and sensor networks (De Silva and Ghrist, 2007).

Several recent attempts have been made, with completely different approaches, to study persistence diagrams from a statistical point of view. One of the first statistical results about persistent homology has been given in a parametric setting in Bubenik and Kim (2007). They show that for data sampled on an hypersphere according to a von-Mises Fisher distribution (among other distributions), the persistence diagrams of the density can be estimated with the parametric rate $n^{-1/2}$. The approach of Mileyko et al. (2011) is completely different, it consists in studying probability measures on the space of persistence diagrams. Bubenik (2015) introduces a functional representation of persistence diagrams, the so-called persistence landscapes, allowing means and variance of persistence diagrams

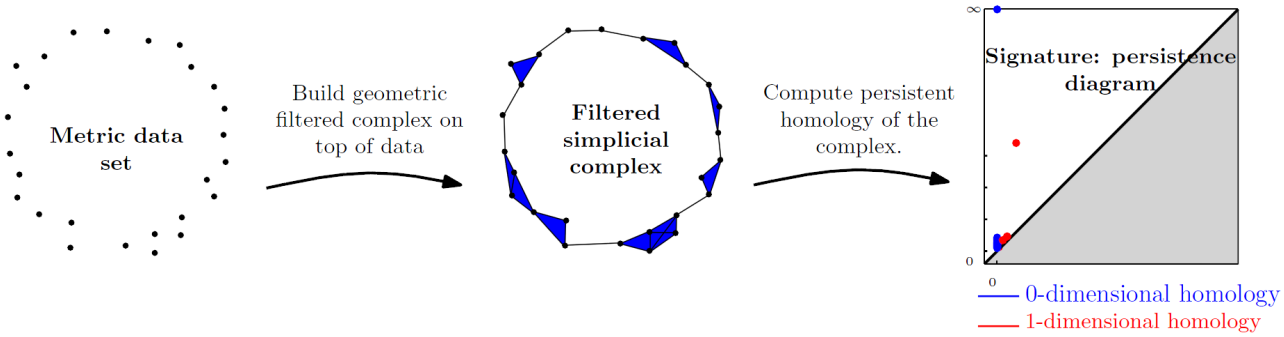


Figure 3.1: A classical pipeline for persistence in TDA.

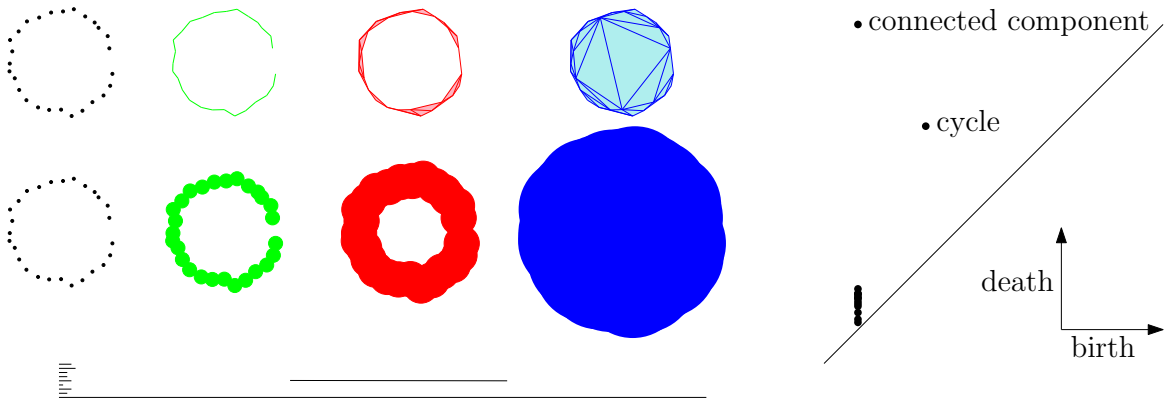


Figure 3.2: Left: an α-complex filtration, the sublevelset filtration of the distance function, and their common persistence barcode. Right: the corresponding persistence diagram.

to be defined, we will follow this approach in Section 3.3.

After recalling the main concepts about persistent homology, we present in this chapter the results of our papers Chazal et al. (2014b) and Chazal et al. (2014c) about rates of convergence of persistence diagram estimation. The third section is about the subsampling methods for persistent homology inference of our paper Chazal et al. (2015a).

3.1 Persistence diagrams and persistence landscapes

Filtrations on metric spaces. The simplicial complexes we consider in this chapter are built on top of metric spaces. As noticed previously in the section 1.2, Čech filtrations and Vietoris-Rips filtrations can be defined in metric spaces. Those filtrations provide a convenient way to study the evolution of the topology of the union of growing balls or the sublevel sets of the distance to a compact, see Fig. 3.2. In the following, the notation $\mathrm{Filt}(\mathbb{X}) := (\mathrm{Filt}_\alpha(\mathbb{X}))_{\alpha\in\mathcal{A}}$ denotes one of these filtrations on a compact set $\mathbb{X}$.

Persistent homology. An extensive presentation of persistence diagrams is available in Chazal et al. (2012). We recall a few definitions and results needed for this chapter and we give the intuition behind persistence. Given a filtration, the topology of $\mathrm{Filt}_\alpha(\mathbb{X})$ changes as α increases: new connected components can appear, existing connected components can merge, cycles and cavities can appear or be filled, etc. Persistent homology tracks these changes, identifies *features* and associates an *interval* or *lifetime* (from α_{birth} to α_{death}) to them. For instance, a connected component is a feature that is born at the smallest α such that the component is present in $\mathrm{Filt}_\alpha(\mathbb{X})$, and dies when it merges with an older connected component. Intuitively, the longer a feature persists, the more relevant it is. Given a filtration as above, we can compute the $\mathbb{Z}_2$-homology and we obtain the homology groups $H(\mathrm{Filt}_\alpha(\mathbb{X})$ at each scale. These groups are also vector spaces $(H(\mathrm{Filt}_\alpha(\mathbb{X})))_{\alpha\in\mathcal{A}}$ and the inclusions $\mathrm{Filt}_\alpha(\mathbb{X}) \subseteq \mathrm{Filt}_\beta(\mathbb{X})$ induce linear maps $H(\mathrm{Filt}_\alpha(\mathbb{X})) \to H(\mathrm{Filt}_\beta(\mathbb{X}))$. In many cases, this sequence can be decomposed as a direct sum of intervals, where an interval is a sequence of the form

$$0 \to \ldots \to 0 \to \mathbb{Z}_2 \to \ldots \to \mathbb{Z}_2 \to 0 \to \ldots \to 0$$

(the linear maps $\mathbb{Z}_2 \to \mathbb{Z}_2$ are all the identity). These intervals can be interpreted as features of the (filtered) complex, such as a connected component or a loop, that appear at parameter α_{birth} in the filtration and disappear at parameter α_{death}. An interval is determined uniquely by these two parameters. A feature, or more precisely its lifetime, can be represented as a segment whose extremities have abscissae α_{birth} and α_{death}; the set of these segments is called the *barcode* of $\mathrm{Filt}(\mathbb{X})$. An interval can also be represented as a point in the plane with coordinates $(\alpha_{\text{birth}}, \alpha_{\text{death}})$, where the x-coordinate indicates the birth time and the y-coordinate the death time (see Fig. 3.2).

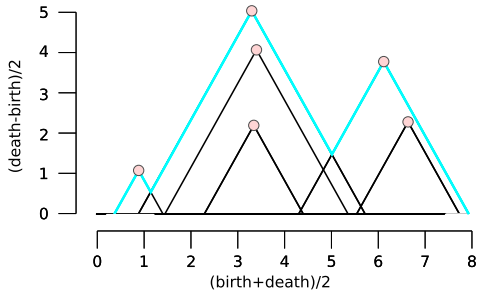


Figure 3.3: We use the rotated axes to represent a persistence diagram Dgm. A feature $(\alpha_{\text{birth}}, \alpha_{\text{death}}) \in \text{Dgm}$ is represented by the point $(\frac{\alpha_{\text{birth}}+\alpha_{\text{death}}}{2}, \frac{\alpha_{\text{death}}-\alpha_{\text{birth}}}{2})$ (pink). In words, the x-coordinate is the average parameter value over which the feature exists, and the y-coordinate is the half-life of the feature. The cyan curve is the landscape $\lambda(1, \cdot)$.

Persistence diagrams. The set of points $(\alpha_{\text{birth}}, \alpha_{\text{death}})$ representing the intervals is called the *persistence diagram* and is denoted $\text{Dgm}(\text{Filt}(\mathbb{X}))$ in the following, see the right picture of Figure 3.2. Note that the diagram is entirely contained in the half-plane above the diagonal Δ defined by $y = x$, since death always occurs after birth. Chazal et al. (2012) shows that this diagram is still well-defined under very weak hypotheses, and in particular $\text{Dgm}(\text{Filt}(\mathbb{X}))$ is well-defined for any compact metric space $\mathbb{X}$ (Chazal et al., 2014d). For technical reasons, the points of the diagonal Δ are considered as part of every persistence diagram, with infinite multiplicity. The most persistent features (supposedly the most important) are those represented by the points furthest from the diagonal in the diagram, whereas points close to the diagonal can be interpreted as (topological) noise.

Bottleneck distance. The space of persistence diagrams is endowed with a metric called the *bottleneck distance* d_b. Given two persistence diagrams, it is defined as the infimum, over all perfect matchings of their points, of the largest L^∞-distance between two matched points, see Fig. 3.4. The presence of the diagonal in all diagrams means we can consider partial matchings of the off-diagonal points, and the remaining points are matched to the diagonal. With more details, given two diagrams Dgm_1 and Dgm_2, we can define a matching m as a subset of $\text{Dgm}_1 \times \text{Dgm}_2$ such that every point of $\text{Dgm}_1 \setminus \Delta$ and $\text{Dgm}_2 \setminus \Delta$ appears exactly once in m. The bottleneck distance is then:

$$\text{d}_\text{b}(\text{Dgm}_1, \text{Dgm}_2) = \inf_{\text{matching } m} \max_{(p,q)\in m} ||q - p||_\infty.$$

Note that points close to the diagonal Δ are easily matched to the diagonal, which fits with their interpretation as irrelevant noise.

Persistence landscapes. The persistence landscape, introduced in Bubenik (2015), is a collection of continuous, piecewise linear functions $\lambda\colon \mathbb{Z}^+ \times \mathbb{R} \to \mathbb{R}$ that summarizes a persistence diagram Dgm, see Fig. 3.3. To define the landscape, consider the set of functions created by tenting each point $p = (x, y) = \left(\frac{\alpha_{\text{birth}}+\alpha_{\text{death}}}{2}, \frac{\alpha_{\text{death}}-\alpha_{\text{birth}}}{2}\right)$ representing a birth-death pair $(\alpha_{\text{birth}}, \alpha_{\text{death}}) \in \text{Dgm}$ as follows:

$$\Lambda_p(t) = \begin{cases} t - x + y & t \in [x - y, x] \\ x + y - t & t \in (x, x + y] \\ 0 & \text{otherwise} \end{cases} = \begin{cases} t - \alpha_{\text{birth}} & t \in [\alpha_{\text{birth}}, \frac{\alpha_{\text{birth}}+\alpha_{\text{death}}}{2}] \\ \alpha_{\text{death}} - t & t \in (\frac{\alpha_{\text{birth}}+\alpha_{\text{death}}}{2}, \alpha_{\text{death}}] \\ 0 & \text{otherwise.} \end{cases}$$

We obtain an arrangement of piecewise linear curves by overlaying the graphs of the functions $\{\Lambda_p\}_p$. The persistence landscape of Dgm is a summary of this arrangement. To avoid minor technical difficulties, we restrict our attention to persistence landscapes for metric spaces $\mathbb{X}$ such that $(\alpha_{\text{birth}}, \alpha_{\text{death}}) \in [0, T] \times [0, T]$ for all $(\alpha_{\text{birth}}, \alpha_{\text{death}}) \in \text{Dgm}(\text{Filt}(\mathbb{X}))$, for some fixed $T > 0$[1]. Formally, the persistence landscape of $\text{Dgm}(\text{Filt}(\mathbb{X}))$ is the collection of functions

$$\lambda_{\text{Dgm}(\text{Filt}(\mathbb{X}))}(k, t) = \operatorname*{kmax}_p \Lambda_p(t), \quad t \in [0, T], k \in \mathbb{N}, \tag{3.1}$$

[1]The point $(0, \infty)$, from zero persistence, is also removed.

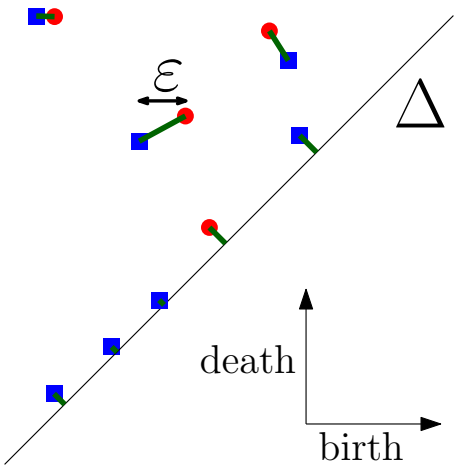


Figure 3.4: Two diagrams at bottleneck distance ε.

where kmax is the kth largest value in the set; in particular, 1max is the usual maximum function. We set $\lambda_{\mathrm{Dgm}(\mathrm{Filt}(\mathbb{X}))}(k,t) = 0$ if the set $\{\Lambda_p(t)\}_p$ contains less than k points. For simplicity of exposition, we use the notation $\lambda_{\mathbb{X}}$ to denote the landscape of $\mathrm{Dgm}(\mathrm{Filt}(\mathbb{X}))$, although the construction depends on the chosen filtration.

Stability. Two compact metric spaces $(\mathbb{X}, \rho)$ and $(\tilde{\mathbb{X}}, \tilde{\rho})$ are *isometric* if there exists a bijection $\Phi : \mathbb{X} \to \tilde{\mathbb{X}}$ that preserves distances. One way to compare two metric spaces is to measure how far these two metric spaces are from being isometric. The corresponding distance is called the *Gromov-Hausdorff distance* $\mathrm{d_{GH}}$ (Burago et al., 2001). Intuitively, it is the infimum of their Hausdorff distance over all possible isometric embeddings of these two spaces into a common metric space. A fundamental property of persistence diagrams, proven in Chazal et al. (2012), is their *stability* with respect to the Gromov-Hausdorff distance, one has

$$\mathrm{d_b}\left(\mathrm{Dgm}(\mathrm{Filt}(\mathbb{X})), \mathrm{Dgm}(\mathrm{Filt}(\tilde{\mathbb{X}}))\right) \leqslant 2\,\mathrm{d_{GH}}\left(\mathbb{X}, \tilde{\mathbb{X}}\right). \tag{3.2}$$

Moreover, if $\mathbb{X}$ and $\tilde{\mathbb{X}}$ are embedded in the same space $(\mathbb{M}, \rho)$ then (3.2) holds for the Hausdorff distance $\mathrm{d_H}$ in place of $\mathrm{d_{GH}}$. From the definition of persistence landscape, we immediately observe that $\lambda(k, \cdot)$ is one-Lipschitz and thus a similar stability is satisfied for the landscapes.

Lemma 1. [Bubenik 2015] *Let $\mathbb{X}$ and $\tilde{\mathbb{X}}$ be two compact sets. For any $t \in \mathbb{R}$ and any $k \in \mathbb{N}$, we have:*
(i) $\lambda_{\mathbb{X}}(k,t) \geqslant \lambda_{\mathbb{X}}(k+1,t) \geqslant 0$.
(ii) $|\lambda_{\mathbb{X}}(k,t) - \lambda_{\tilde{\mathbb{X}}}(k,t)| \leqslant \mathrm{d_b}(\mathrm{Dgm}(\mathrm{Filt}(\mathbb{X})), \mathrm{Dgm}(\mathrm{Filt}(\tilde{\mathbb{X}})))$.

3.2 Estimation of persistent diagrams on metric spaces

Assume that we observe n points $X_1 \ldots, X_n$ in a metric space $(\mathbb{M}, \rho)$ drawn i.i.d. from some unknown measure μ whose support is a compact set denoted $\mathbb{X}_\mu$. The Gromov-Hausdorff distance allows us to compare $\mathbb{X}_\mu$ with compact metric spaces not necessarily embedded in $\mathbb{M}$. In the following, an *estimator* $\widehat{\mathbb{X}}$ of $\mathbb{X}_\mu$ is a function of $X_1 \ldots, X_n$ that takes values in the set of compact metric spaces and which is measurable for the Borel algebra induced by $\mathrm{d_{GH}}$.

Let $\mathrm{Filt}(\mathbb{X}_\mu)$ and $\mathrm{Filt}(\widehat{\mathbb{X}})$ be two filtrations defined on $\mathbb{X}_\mu$ and $\widehat{\mathbb{X}}$. Starting from (3.2) our strategy consists in estimating the support $\mathbb{X}_\mu$ with respect to the $\mathrm{d_{GH}}$ distance. Note that this general strategy of estimating $\mathbb{X}_\mu$ in $\mathcal{K}$ is not only of theoretical interest. Indeed, in some cases the space $\mathbb{M}$ is unknown and the observations $X_1 \ldots, X_n$ are just known through their pairwise distances $\rho(X_i, X_j)$, $i, j = 1, \cdots, n$. The use of the Gromov-Hausdorff distance then allows us to consider this set of observations as an abstract metric space of cardinality n, independently of the way it is embedded in $\mathbb{M}$. This general framework includes the more standard approach consisting in estimating the support with respect to the Hausdorff distance by restraining the values of $\widehat{\mathbb{X}}$ to the compact sets included in $\mathbb{M}$.

Let $\mathbb{X}_n := \{X_1, \ldots, X_n\}$ be a set of independent observations endowed with the restriction of the distance ρ to this set. This finite metric space is a natural estimator of the support $\mathbb{X}_\mu$. In several contexts discussed in the following, $\mathbb{X}_n$ shows optimal rates of convergence to $\mathbb{X}_\mu$ with respect to the Hausdorff distance.

The rate of convergence of $\mathbb{X}_n$ in Gromov-Hausdorff distance is obtained under the (a, b)-standard assumption: for some constants $a, b > 0$: for any $x \in \mathbb{X}_\mu$ and any $r > 0$,

$$\mu(B(x, r)) \geqslant \min(ar^b, 1). \tag{3.3}$$

This assumption has been widely used in the literature of set estimation under Hausdorff distance (Cuevas and Rodríguez-Casal, 2004; Singh et al., 2009).

Theorem 3. [Chazal et al. 2014b] *Assume that the probability measure μ on $\mathbb{M}$ satisfies the (a, b)-standard assumption, then for any $\varepsilon > 0$:*

$$\mathbb{P}\left(\mathrm{d_b}\left(\mathrm{Dgm}(\mathrm{Filt}(\mathbb{X}_\mu)), \mathrm{Dgm}(\mathrm{Filt}(\mathbb{X}_n))\right) > \varepsilon\right) \leqslant \min\left(\frac{2^b}{a\varepsilon^b} \exp(-na\varepsilon^b), 1\right). \tag{3.4}$$

Moreover,

$$\limsup_{n\to\infty} \left(\frac{n}{\log n}\right)^{1/b} \mathrm{d_b}(\mathrm{Dgm}(\mathrm{Filt}(\mathbb{X}_\mu)), \mathrm{Dgm}(\mathrm{Filt}(\mathbb{X}_n))) \leqslant C_1$$

almost surely, and

$$\mathbb{P}\left(\mathrm{d_b}\left(\mathrm{Dgm}(\mathrm{Filt}(\mathbb{X}_\mu)), \mathrm{Dgm}(\mathrm{Filt}(\mathbb{X}_n))\right) \leqslant C_2 \left(\frac{\log n}{n}\right)^{1/b}\right)$$

converges to 1 when $n \to \infty$, where C_1 and C_2 only depend on a and b.

Let $\mathcal{P} = \mathcal{P}(a, b, \mathbb{M})$ be the set of all the probability measures on the metric space $(\mathbb{M}, \rho)$ satisfying the (a, b)-standard assumption on $\mathbb{M}$:

$$\mathcal{P} := \left\{\mu \text{ on } \mathbb{M} \mid \mathbb{X}_\mu \text{ is compact and } \forall x \in \mathbb{X}_\mu, \forall r > 0,\ \mu\left(B(x, r)\right) \geqslant \min\left(1, ar^b\right)\right\}. \tag{3.5}$$

The next theorem gives upper and lower bounds for the rate of convergence of persistence diagrams. The upper bound is a consequence of Theorem 3, while the lower bound is established using Le Cam's lemma.

Theorem 4. [Chazal et al. 2014b] *For some positive constants a and b,*

$$\sup_{\mu\in\mathcal{P}} \mathbb{E}\left[\mathrm{d_b}(\mathrm{Dgm}(\mathrm{Filt}(\mathbb{X}_\mu)), \mathrm{Dgm}(\mathrm{Filt}(\mathbb{X}_n)))\right] \leqslant C \left(\frac{\log n}{n}\right)^{1/b}$$

where the constant C only depends on a and b (not on $\mathbb{M}$). Assume moreover that there exists a non isolated point x in $\mathbb{M}$ and consider any sequence $(x_n) \in (\mathbb{M}\backslash\{x\})^{\mathbb{N}}$ such that $\rho(x, x_n) \leqslant (an)^{-1/b}$. Then for any estimator $\widehat{\mathrm{Dgm}}_n$ of $\mathrm{Dgm}(\mathrm{Filt}(\mathbb{X}_\mu))$:

$$\liminf_{n\to\infty} \rho(x, x_n)^{-1} \sup_{\mu\in\mathcal{P}} \mathbb{E}\left[\mathrm{d_b}(\mathrm{Dgm}(\mathrm{Filt}(\mathbb{X}_\mu)), \widehat{\mathrm{Dgm}}_n)\right] \geqslant C'$$

where C' is an absolute constant.

Consequently, the estimator $\mathrm{Dgm}(\mathrm{Filt}(\mathbb{X}_n))$ is minimax optimal on the space $\mathcal{P}(a, b, \mathbb{M})$ up to a logarithmic term as soon as we can find a non-isolated point in $\mathbb{M}$ and a sequence (x_n) in $\mathbb{M}$ such that $\rho(x_n, x) \sim (an)^{-1/b}$. This is obviously the case for the Euclidean space $\mathbb{R}^d$. One classical method to obtain tight lower bounds with sup norm metrics is applying a Fano's strategy based on several hypotheses (see for instance Tsybakov, 2009, Chapter 2). Applying this method is more difficult than it seems in our context. Indeed, the bottleneck distance makes tricky the construction of multiple hypotheses. However, in specific cases, we can obtain the matching lower bound with a more direct proof.

Theorem 5. [Chazal et al. 2014b] *Consider* $(\frac{1}{2}, 1)$*-standard measures on the unit segment* $[0, 1]$*. For any estimator* $\widehat{\mathrm{Dgm}}_n$ *of* $\mathrm{Dgm}(\mathrm{Filt}(\mathbb{X}_\mu))$*:*

$$\liminf_{n\to\infty} \sup_{\mu\in\mathcal{P}(\frac{1}{2},1,[0,1])} \frac{n}{\log n}\mathbb{E}\left[\mathrm{d_b}(\mathrm{Dgm}(\mathrm{Filt}(\mathbb{X}_\mu)), \widehat{\mathrm{Dgm}}_n)\right] \geqslant C.$$

where C *is an absolute constant.*

It should be straightforward to extend this to measures on the cube $[0, 1]^b$, as long as b is an integer, with a lower-bound of $C_b(\frac{\log n}{n})^{1/b}$. Note that this bound applies to the homology of dimension b. It is possible that lower-dimensional homology may be easier to estimate.

Confidence regions. Theorem 3 can also be used to find confidence sets for persistence diagrams. Such confidence sets depend on a and b which may be unknown and whose estimation is a difficult problem. Alternative solutions have been proposed in Fasy et al. (2014) using subsampling methods and kernel estimators among other approaches, in the specific context of smooth manifolds of an Euclidean space. Note that both Fasy et al. (2014) and our work start from the observation that persistence diagram inference is strongly connected to the better known problem of support estimation.

Persistence diagram estimation for nonsingular measures in $\mathbb{R}^k$. Assume that μ is a measure on $\mathbb{R}^k$ with density f with respect to Lebesgue. Following Singh et al. (2009), assume (among other assumptions) that in the neighborhood of the boundary $\partial\mathbb{X}_\mu$ of $\mathbb{X}_\mu$, $f(x) \geqslant C\, d(x, \partial\mathbb{X}_\mu)^\alpha$. We prove that $\mathrm{Dgm}(\mathrm{Filt}(\mathbb{X}_n))$ converges in expectation to $\mathrm{Dgm}(\mathrm{Filt}(\mathbb{X}_\mu))$ with a rate upper bounded by $(\log n/n)^{1/(k+\alpha)}$. Moreover, it can be shown that this rate is minimax over a convenient family of densities with respect to the Lebesgue measure on $\mathbb{R}^k$.

Persistence diagram estimation for singular measures in $\mathbb{R}^D$. Let μ be a measure supported on a smooth submanifold of $\mathbb{R}^D$ with positive reach. Assume that μ has a density with respect to the k-dimensional volume measure on $\mathbb{X}_\mu$, which is lower and upper bounded on $\mathbb{X}_\mu$. From Genovese et al. (2012), we obtain that $\mathrm{Dgm}(\mathrm{Filt}(\mathbb{X}_n))$ converges in expectation to $\mathrm{Dgm}(\mathrm{Filt}(\mathbb{X}_\mu))$ with a rate upper bounded by $(\frac{\log n}{n})^{1/k}$ both for support and persistence diagram estimation. Nevertheless, this rate is not minimax optimal for support estimation, as shown by Theorem 2 in Genovese et al. (2012). The correct minimax rate is actually $n^{-2/k}$ for both estimation problems.

Additive noise. Consider the convolution model where the observations satisfy $Y_i = X_i + \varepsilon_i$ where $X_1, \dots X_n$ are sampled according to a measure μ as in the previous paragraph and where $\varepsilon_1, \dots, \varepsilon_n$ are i.i.d. standard Gaussian random variables. We deduce from the results of Genovese et al. (2012) that the minimax convergence rates for the persistence diagram estimation in this context is upper bounded by some rate of the order of $(\log n)^{-1/2}$. However, giving a tight lower bound for this problem appears to be more difficult than for the support estimation problem.

Persistence landscapes. According to the stability results given in Lemma 1, upper bounds on the rates of convergence of the persistence landscapes directly derive from our results. A complete minimax description of the problem would also require to prove the corresponding lower bounds.

3.3 Subsampling methods for persistent homology

The time and space complexity of persistent homology algorithms is one of the main obstacles in applying TDA techniques to high-dimensional problems. To overcome the problem of computational costs, we propose in Chazal et al. (2015a) the following strategy: given a large point cloud, take several subsamples, compute the landscape for each subsample, and then combine the information. Indeed, contrary to persistence diagrams, persistence landscape can be averaged in a straightforward way.

As in the previous section, a probability measure μ is defined on a metric space $(\mathbb{M}, \rho)$ and the support of μ is a compact set $\mathbb{X}_\mu$. In all the section it is assumed that the diameter of $\mathbb{M}$ is finite and

upper bounded by $\frac{T}{2}$, where T is the same constant as in the definition of persistence landscapes in Section 3.1. For ease of exposition, we focus on the case $k = 1$, and set $\lambda(t) = \lambda(1, t)$. However, the results we present in this section hold for $k > 1$.

3.3.1 The multiple samples approach

For any positive integer m, let $X = \{x_1, \cdots, x_m\} \subset \mathbb{X}_\mu$ be a sample of m points from μ. The corresponding persistence landscape is λ_X and we denote by Ψ_μ^m the measure induced by $\mu^{\otimes m}$ on the space of persistence landscapes. Note that the persistence landscape λ_X can be seen as a single draw from the measure Ψ_μ^m. We consider the point-wise expectations of the (random) persistence landscape under this measure: $\mathbb{E}_{\Psi_\mu^m}[\lambda_X(t)], t \in [0, T]$. The average landscape $\mathbb{E}_{\Psi_\mu^m}[\lambda_X]$ has a natural empirical counterpart, which can be used as its unbiased estimator. Let $S_1^m, \ldots, S_\ell^m$ be ℓ independent samples of size m from $\mu^{\otimes m}$. We define the empirical average landscape as

$$\overline{\lambda_\ell^m}(t) = \frac{1}{b} \sum_{i=1}^{b} \lambda_{S_i^m}(t), \qquad \text{for all } t \in [0, T], \tag{3.6}$$

and propose to use $\overline{\lambda_\ell^m}$ to estimate $\lambda_{\mathbb{X}_\mu}$.

In addition to the average, we also consider using the *closest sample* to $\mathbb{X}_\mu$ in Hausdorff distance. The closest sample method consists in choosing a sample of m points of $\mathbb{X}_\mu$, as close as possible to $\mathbb{X}_\mu$, and then use this sample to build a landscape that approximates $\lambda_{\mathbb{X}_\mu}$. Let $S_1^m, \ldots, S_\ell^m$ be ℓ independent samples of size m from $\mu^{\otimes m}$. The closest sample is

$$\widehat{C}_\ell^m = \arg \min_{S \in \{S_1^m, \ldots, S_\ell^m\}} \mathrm{d_H}(S, X_\mu) \tag{3.7}$$

and the corresponding landscape function is $\hat{\lambda}_\ell^m = \lambda_{\widehat{C}_\ell^m}$. Of course, the method requires the support of μ to be a known quantity.

Computing the persistent homology of $\mathbb{X}_n$ is $O(\exp(n))$, whereas computing the average landscape is $O(b \exp(m))$ and the persistent homology of the closest sample is $O(bmn + \exp(m))$.

Remark 1. *The general framework described above is valid for the case in which μ is a discrete measure with support $\mathbb{X}_N = \{x_1, \ldots, x_N\} \subset \mathbb{R}^D$. This framework can be very common in practice, when a continuous (but unknown measure) is approximated by a discrete uniform measure μ_N on $\mathbb{X}_N$.*

3.3.2 Stability of the average landscape

We show below that the average landscape $\mathbb{E}_{\Psi_\mu^m}[\lambda_X]$ is an interesting quantity on its own, since it carries some stable topological information about the underlying measure μ, from which the data are generated. In particular, we will compare the average landscapes corresponding to two measures that are close to each other in the Wasserstein metric (see for instance Rachev and Rüschendorf, 1998; Villani, 2008).

Definition 7. *The pth Wasserstein distance between two measures μ, ν defined on $(\mathbb{M}, \rho)$ is*

$$W_{\rho,p}(\mu, \nu) = \left(\inf_{\Pi} \int_{\mathbb{M} \times \mathbb{M}} [\rho(x, y)]^p d\Pi(x, y) \right)^{\frac{1}{p}},$$

where the infimum is taken over all measures on $\mathbb{M} \times \mathbb{M}$ with marginals μ and ν.

First, we show that the average behavior of the landscapes of sets of m points sampled according to any measure μ is stable with respect to the Wasserstein distance.

Theorem 6. [Chazal et al. 2015a] *Let $X \sim \mu^{\otimes m}$ and $Y \sim \nu^{\otimes m}$, where μ and ν are two probability measures on $\mathbb{M}$. For any $p \geqslant 1$ we have*

$$\left\| \mathbb{E}_{\Psi_\mu^m}[\lambda_X] - \mathbb{E}_{\Psi_\nu^m}[\lambda_Y] \right\|_\infty \leqslant 2\, m^{\frac{1}{p}} W_{\rho,p}(\mu, \nu).$$

Remark 2. *For measures that are not defined on the same metric space, the inequality of Theorem 6 can be extended to Gromov-Wasserstein metric:* $\left\|\mathbb{E}_{\Psi_\mu^m}[\lambda_X] - \mathbb{E}_{\Psi_\nu^m}[\lambda_Y]\right\|_\infty \leqslant 2m^{\frac{1}{p}} GW_{\rho,p}(\mu,\nu)$.

The result of Theorem 6 is useful for two reasons. First, it tells us that for a fixed m, the expected "topological behavior" of a set of m points carries some stable information about the underlying measure from which the data are generated. Second, it provides a lower bound for the Wasserstein distance between two measures, based on the topological signature of samples of m points.

The dependence on m of the upper bound of Theorem 6 seems to be necessary in this setting: intuitively, when m grows, the samples of m points converge to the support of μ and ν w.r.t. the Hausdorff distance. Therefore the expected landscapes should converge to the landscapes of the support of the measures. But, in general, two measures that are close in the Wasserstein metric can have support that have very different and unrelated topologies. Indeed, a similar dependence was also obtained in Blumberg et al. (2014) when considering the Gromov-Prohorov metric.

Note that in Theorem 6 we do not make any assumption on the measures μ and ν. If we assume that they both μ and ν satisfy the (a,b,r_0)-standard assumption (defined below) we can provide a different bound on the difference of the expected landscapes, based on the Hausdorff distance between the support of the two measures. We say that μ satisfies the (a,b,r_0)-standard assumption if there exist positive constants a, b and $r_0 \geqslant 0$ such that

$$\forall r > r_0,\ \forall x \in \mathbb{X}_\mu\ ,\ \mu(B(x,r)) \geqslant 1 \wedge a r^b. \tag{3.8}$$

The case $r_0 = 0$ corresponds to the (a,b)-standard assumption (3.3). We use the generalized version with $r_0 > 0$ to take into account the case in which μ is a discrete measure, see Remark 1, in which case r_0 depends on N.

Theorem 7. [Chazal et al. 2015a] *Let $X \sim \mu^{\otimes m}$ and $Y \sim \nu^{\otimes m}$, where μ and ν are two probability measures on $\mathbb{M}$ which both satisfy the (a,b,r_0)-standard assumption on their support. Define $r_m = 2\left(\frac{\log m}{am}\right)^{1/b}$. Then*

$$\|\mathbb{E}_{\Psi_\mu^m}(\lambda_X) - \mathbb{E}_{\Psi_\nu^m}(\lambda_Y)\|_\infty \leqslant 2\,\mathrm{d}_H(\mathbb{X}_\mu,\mathbb{X}_\nu) + 4r_0 + 4r_m \mathbf{1}_{(r_0,\infty)}(r_m)\ +\ 4\,C_1(a,b)\,r_m\,\frac{1}{(\log m)^2},$$

where $C_1(a,b)$ is a constant depending on a and b.

3.3.3 Risk analysis

We now study the performances of $\overline{\lambda_\ell^m}$ and $\hat{\lambda}_\ell^m$, as estimators of $\lambda_{\mathbb{X}_\mu}$. We start by decomposing the ℓ_∞-risk of the average landscape as follows. Set $\lambda_1 = \lambda_{S_1^m}$, with S_1^m a sample of size m from μ. Then,

$$\mathbb{E}\left\|\lambda_{\mathbb{X}_\mu} - \overline{\lambda_\ell^m}\right\|_\infty \leqslant \left\|\lambda_{\mathbb{X}_\mu} - \mathbb{E}\lambda_1\right\|_\infty + \mathbb{E}\left\|\overline{\lambda_\ell^m} - \mathbb{E}\lambda_1\right\|_\infty, \tag{3.9}$$

where the expectation of $\overline{\lambda_\ell^m}$ is w.r.t. $(\Psi_\mu^m)^{\otimes b}$ and the expectation of λ_1 is w.r.t. Ψ_μ^m. For the bias term $\left\|\lambda_{\mathbb{X}_\mu} - \mathbb{E}\lambda_1\right\|_\infty$ we use the stability property to go back into $\mathbb{R}^d$:

$$\left\|\lambda_{\mathbb{X}_\mu} - \mathbb{E}\lambda_1\right\|_\infty \leqslant \mathbb{E}_{\Psi_\mu^m}\left\|\lambda_{\mathbb{X}_\mu} - \lambda_1\right\|_\infty \leqslant 2\mathbb{E}_{\mu^{\otimes m}}\,\mathrm{d_H}(\mathbb{X}_\mu, X), \tag{3.10}$$

where X is a sample of size m from μ. Note that, if calculating $\mathrm{d_H}(\mathbb{X}_\mu, X)$ is computationally feasible, then, in practice, $\mathbb{E}_{\mu^{\otimes m}}\,\mathrm{d_H}(\mathbb{X}_\mu, X)$ can be approximated by the average of a large number B of values of $\mathrm{d_H}(\mathbb{X}_\mu, X)$, for B different draws of subsamples $X \sim \mu^{\otimes m}$. To give an explicit bound on the bias, we assume that μ satisfies the (a,b,r_0)-standard assumption.

Theorem 8. [Chazal et al. 2015a] *Let $r_m = 2\left(\frac{\log m}{am}\right)^{1/b}$. If μ satisfies the (a,b,r_0)-standard assumption, then*

$$\left\|\lambda_{\mathbb{X}_\mu} - \mathbb{E}\lambda_1\right\|_\infty \leqslant 2r_0\ +\ 2r_m\,\mathbf{1}_{(r_0,\infty)}(r_m)\ + 2C_1(a,b)\,r_m\,\frac{1}{(\log m)^2},$$

where $C_1(a,b)$ is a constant that depends on a and b.

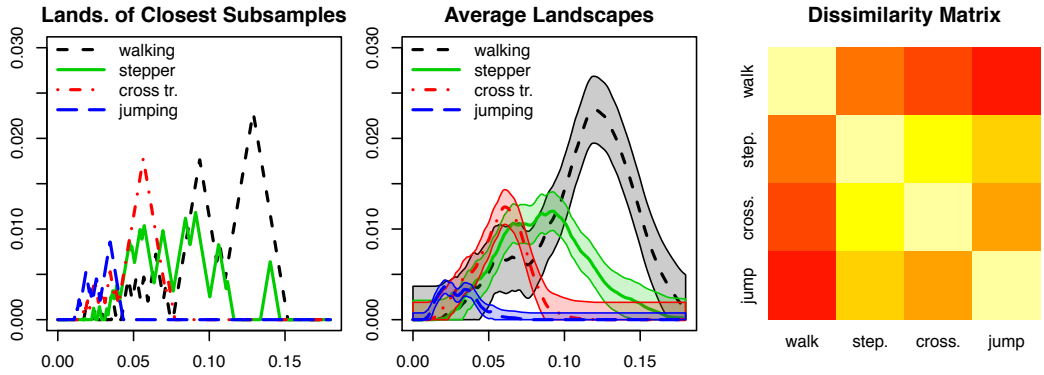


Figure 3.5: Subsampling methods applied to magnetometer data. For $\ell = 80$ subsamples of size $m = 200$, for each activity, we constructed the landscapes of the closest subsample (left), the average landscape with 95% confidence band (middle) and the dissimilarity matrix of the pairwise ℓ_∞ distance between average landscapes.

The analysis of the variance term can be found in Chazal et al. (2014e), it is of the order of $1/\sqrt{\ell}$. Therefore, if r_0 is negligible, we see that ℓ should be taken of the order of $(m/\log m)^{2/b}$.

We now turn to the closest sample estimator $\hat{\lambda}_\ell$ and investigate its ℓ_∞ risk $\mathbb{E}\left[\|\lambda_{\mathbb{X}_\mu} - \hat{\lambda}_\ell^m\|_\infty\right]$, where the expectation is with respect to $(\Psi_\mu^m)^{\otimes \ell}$. As before, our analysis relies on the stability property

$$\mathbb{E}\left[\|\lambda_{\mathbb{X}_\mu} - \lambda_\ell^{\hat{m}}\|_\infty\right] \leqslant 2\mathbb{E}\left[\mathrm{d_H}(\mathbb{X}_\mu, \widehat{C_\ell^m})\right],$$

where the second expectation is with respect to $(\mu^{\otimes m})^{\otimes \ell}$.

Theorem 9. [Chazal et al. 2015a] *Let* $r_m = 2\left(\frac{\log(2^b m)}{am}\right)^{\frac{1}{b}}$. *If* $\mu \in \mathcal{P}(\mathbb{X})$ *satisfies the* (a, b, r_0)*-standard assumption, then*

$$\mathbb{E}\left[\|\lambda_{\mathbb{X}_\mu} - \lambda_\ell^{\hat{m}}\|_\infty\right] \leqslant 2r_0 \;+\; 2r_m \mathbf{1}_{(r_0,\infty)}(r_m) \;+ 2C_2(a,b)\, r_m \frac{1}{\ell\,[\log(2^b m)]^{\ell+1}},$$

where $C_2(a, b)$ *is a constant that depends on* a *and* b.

The risk of the closest subsample method can in principle be smaller than the average landscape method. In particular, if μ is the discrete uniform measure on a point cloud of size N, sampled from a measure satisfying the $(a, b, 0)$-standard assumption, then it can be shown that r_0 is of the order of $(\frac{\log N}{N})^{1/b}$. When r_0 is negligible, the rates of theorems 8 and 9 are comparable, both of the order of $O(\frac{\log m}{m})^{1/b}$.

3.4 Experiments

In Chazal et al. (2015a), we apply the method to the problem of distinguishing human activities performed while wearing inertial and magnetic sensor units. The dataset is publicly available at the UCI Machine Learning Repository[2] and is described in Barshan and Yüksek (2013), where it is used to classify 19 activities performed by eight people wearing sensor units on the chest, arms, and legs. For ease of illustration, we report here the results on four activities (walking, stepper, cross trainer, jumping) performed by a single person. We use the data from the magnetometer of a single sensor (left leg), which measures the direction of the magnetic field in the space at a frequency of 25Hz. For each activity there are 7,500 consecutive measurements that we treat as a 3D point cloud in the Euclidean space. For $\ell = 80$ times, we subsample $m = 200$ points from the point cloud of each activity, then construct the landscapes of the closest subsamples, the average landscapes (dimension one), and the dissimilarity matrix based on the ℓ_∞ distances of the average landscapes, see Fig. 3.5. To the

[2] http://archive.ics.uci.edu/ml/datasets/Daily+and+ Sports+Activities

best of our knowledge, persistent homology has never been used to study data from accelerometers or magnetometers before. A remarkable advantage is that the methods of persistent homology are insensitive to the orientation of the input data, as opposed to other methods that require the exact calibration of the sensor units; see, for example, Altun et al. (2010) and Barshan and Yüksek (2013).

3.5 Discussion and directions for future research

Persistent homology for supervised problems. Persistent homology has been successfully applied for many real-life problems but as far as we know, it has been essentially used with for exploring data (Carlsson et al., 2012). In many situations as for the Magnetometer Data of last section, we believe that the topological information that can be extracted from the point clouds may help to predict an outcome variable. We are currently working on (large) real life datasets in order to demonstrate the interest of persistent homology methods for supervised problems.

Persistent homology of the kernel estimator of the density. Persistent homology can be used to study the homology of the level sets of a given function. For instance, Fasy et al. (2014) study the super-level sets of a kernel density estimator. This approach is motivated by the fact that, in favorable cases, some level sets of the density are homotopy equivalent to the support $\mathbb{X}$ of the distribution (without any noise). This implies that we could estimate the homology of $\mathbb{X}$ from density estimators. But as far as we know, the complete risk analysis of the kernel density estimator, when this last is used to estimate the persistence homology of the support of the distribution, has not been proposed so far.

Tuning parameters for persistent homology. As already noticed in the previous chapter, an unsolved problem in topological inference is tuning parameters. In the context of the previous paragraph for instance, choosing the bandwidth is a tricky question. We know from Fasy et al. (2014) that this choice depends on the geometry (the reach) of the support but of course these quantity are unknown in practice. Guibas et al. (2013) has suggested tracking the evolution of the persistence of the homological features as the tuning parameter varies. This idea has been formalized into the the so-called *Maximum Significant Topological Signal Strength* in Wasserman (2014) and Chazal et al. (2014a). We intend to prove that this criterion leads to efficient choice, at least in simple frameworks. We also would like to apply this idea for the subsampling methods of Section 3.3 in order to select m in an efficient way. Indeed, we have illustrated with some examples in Chazal et al. (2015a) that the average method is robust to outliers as long as m is not too large.

Robustness to noise is indeed a serious problem for topological inference. In the next chapter, we present our contributions on this topic using the distance to measure function, an alternative distance function having the great advantage of being robust to noise.

Chapter 4

Robust topological data analysis with the distance to measure

.

It is well known that TDA methods may fail completely in the presence of outliers. Indeed, adding even a single outlier to the point cloud can change the distance function dramatically, see Fig. 4.1 for an illustration. To answer this drawback, Chazal et al. (2011b) have introduced an alternative distance function which is robust to noise, the *distance-to-a-measure* (DTM). This chapter presents our contributions on the statistical analysis of the DTM.

Section 4.1 introduces the DTM. Sections 4.2 and 4.3 are about the convergence of the DTM and also about the bootstrap for the DTM. These results come from our papers Chazal et al. (2014a) and Chazal et al. (2015b). Section 4.4 is about deconvolution methods under Wasserstein metrics, a problem which is related to the estimation of the DTM in convolution models. This last section presents the results of our papers Caillerie et al. (2011); Dedecker and Michel (2013); Dedecker et al. (2015).

4.1 The distance to measure

Given a probability distribution P in $\mathbb{R}^d$ and a real parameter $0 \leqslant u \leqslant 1$, Chazal et al. (2011b) have generalized the notion of distance to the support of P by the function

$$\delta_{P,u} : x \in \mathbb{R}^d \mapsto \inf\{t > 0\,;\, P(B(x,t)) \geqslant u\},$$

where $B(x,t)$ is the closed Euclidean ball of center x and radius t. To avoid issues due to discontinuities of the map $P \to \delta_{P,u}$, the distance to measure function with parameter $m \in [0,1]$ and power $r \geqslant 1$ is defined by

$$d_{P,m,r}(x) : x \in \mathbb{R}^d \mapsto \left(\frac{1}{m}\int_0^m \delta_{P,u}^r(x)du\right)^{1/r}. \tag{4.1}$$

A nice property of the DTM proved in Chazal et al. (2011b) is its stability with respect to perturbations of P in the Wasserstein metric. More precisely, the map $P \to d_{P,m,r}$ is $m^{-\frac{1}{r}}$-Lipschitz, i.e. if P and $\tilde{P}$ are two probability distributions on $\mathbb{R}^d$, then

$$\|d_{P,m,r} - d_{\tilde{P},m,r}\|_\infty \leqslant m^{-\frac{1}{r}} W_r(P,\tilde{P}) \tag{4.2}$$

where W_r is the Wasserstein distance for the Euclidean metric on $\mathbb{R}^d$, with power r (see Definition 7 in the previous chapter). This property implies that the DTM associated to close distributions in the Wasserstein metric have close sublevel sets. Moreover, when $r = 2$, the function $d^2_{P,m,2}$ is semiconcave ensuring strong regularity properties on the geometry of its sublevel sets. Using these properties, Chazal et al. (2011b) (Section 4) show that, under general assumptions, if $\tilde{P}$ is a probability distribution approximating P, then the sublevel sets of $d_{\tilde{P},m,2}$ provide a topologically correct approximation of the support of P.

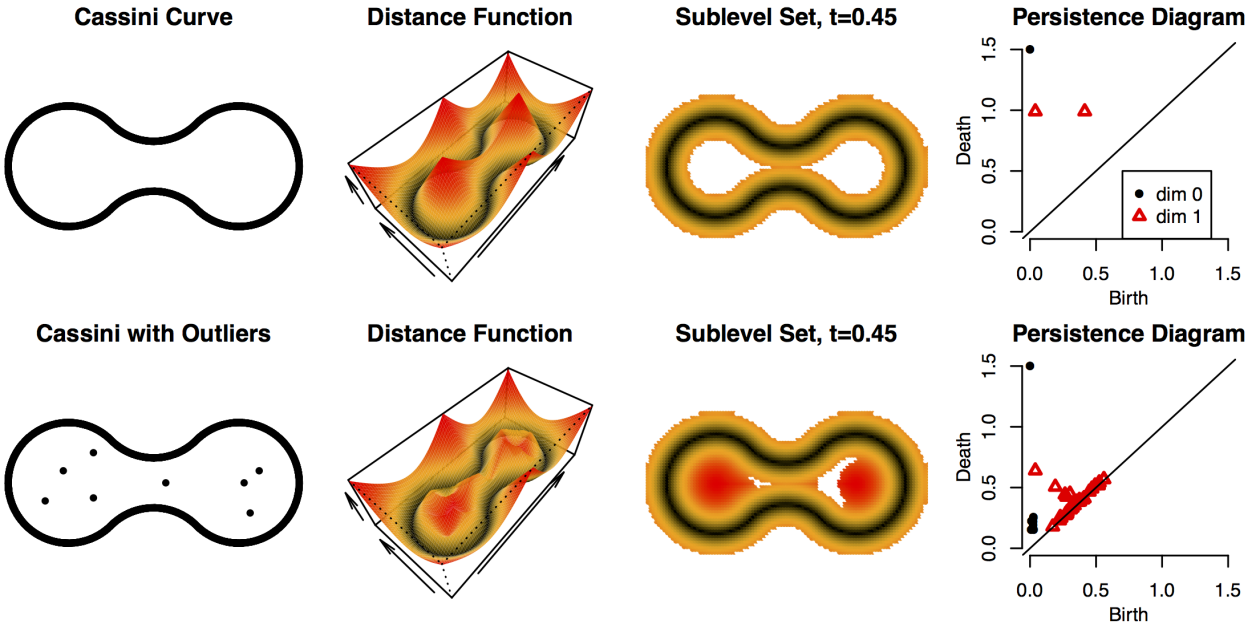


Figure 4.1: Top: Cassini curve $\mathbb{X}$, the distance function $d_{\mathbb{X}}$, a typical sublevel set $\{x : \ d_{\mathbb{X}}(x) \leqslant t\}$ and the resulting persistence diagram. Bottom: the effect of adding a few outliers. The distance function and persistence diagram are dramatically different.

The introduction of DTM has motivated further works and applications in various directions such as topological data analysis (Buchet et al., 2015a), GPS traces analysis (Chazal et al., 2011a), density estimation (Biau et al., 2011), or clustering (Chazal et al., 2013) just to name a few. Approximations, generalizations and variants of the DTM have also been considered in (Guibas et al., 2013; Phillips et al., 2014; Buchet et al., 2015b).

In practice, the measure P is usually only known through a finite set of observations $\mathbb{X}_n = \{X_1, \ldots, X_n\}$ sampled from P, raising the question of the approximation of the DTM. A natural idea to estimate the DTM from $\mathbb{X}_n$ is to plug the empirical measure P_n instead of P in the definition of the DTM. This "plug-in strategy" corresponds to computing the distance to the empirical measure (DTEM). For $m = \frac{k}{n}$, the DTEM satisfies

$$d^r_{P_n,k/n,r}(x) := \frac{1}{k}\sum_{j=1}^{k} \|x - \mathbb{X}_n\|^r_{(j)}\,,$$

where $\|x - \mathbb{X}_n\|_{(j)}$ denotes the distance between x and its j-th neighbor in $\{X_1, \ldots, X_n\}$. This quantity can be easily computed in practice since it only requires the distances between x and the sample points.

Let $F^{-1}_{x,r}$ be the distribution function of the push forward probability measure of P by the function $\|x - \cdot\|^r$ and let $F_{x,n,r}$ be the empirical distribution function of the observed distances (to the power r): $\|x - X_1\|^r, \ \ldots, \ \|x - X_n\|^r$. The corresponding quantile functions are denoted $F^{-1}_{x,r}$ and $F^{-1}_{x,n,r}$. Sections 4.2 and 4.3 study the quantity

$$\begin{aligned} \Delta_{n,m,r}(x) &:= d^r_{P_n,m,r}(x) - d^r_{P,m,r}(x) && (4.3)\\ &= \frac{1}{m}\int_0^m \left\{F^{-1}_{x,n,r}(u) - F^{-1}_{x,r}(u)\right\} du. && (4.4)\end{aligned}$$

We fix $r \geqslant 1$ and we henceforth write F_x for $F_{x,r}$ to facilitate the reading. In the same way we will use the notation F^{-1}_x, $F_{x,n}$, $F^{-1}_{x,n}$ and $\Delta_{P,m}$ since there is no ambiguity on the power term r.

4.2 Rates of convergence of the DTEM

Upper bounds on the rate of convergence of the DTEM can be deduced from the Wasserstein stability (4.2) of the DTM together with known results about the convergence of the empirical measure under Wasserstein metrics: using recent results of Fournier and Guillin (2013) or from Dereich et al. (2013), we find that $\mathbb{E}\|\Delta_{n,\frac{k}{n},r}\|_\infty \lesssim n^{-1/d}$. However the Wasserstein stability results are not tight enough to provide the correct rates of convergence of the DTM. To obtain better upper bounds on the rate of convergence of the DTEM, we directly control the fluctuations of $\Delta_{n,\frac{k}{n},r}$ by considering a supremum over the underlying empirical process.

4.2.1 Local analysis of the DTEM in the bounded case

We first consider the behavior of the distance to the empirical measure when the observations $X_1, \dots, X_n$ are sampled from a distribution P with compact support in $\mathbb{R}^d$. Let F_x^{-1} be the quantile function of $\|x - X_1\|^r$ for some observation point $x \in \mathbb{R}^d$. We introduce the modulus of continuity $\tilde{\omega}_x$ of F_x^{-1} (possibly infinite) which is defined for any $v \in (0,1]$ by

$$\tilde{\omega}_x(v) := \sup_{(u,u')\in[0,1]^2,\, u\neq u',\, \|u-u'\|\leqslant v} |F_x^{-1}(u) - F_x^{-1}(u')|.$$

In the following we consider a (continuous) upper bounds on the modulus of continuity, that is a non negative function ω_x on $(0,1]$ such that $\omega_x(v) \geqslant \tilde{\omega}(v)$ for any $v \in (0,1]$. A modulus of continuity being a non decreasing function, we will assume that such an upper bound ω_x is non decreasing on $(0,1]$.

Theorem 10. [Chazal et al. 2015b] *Let x be a fixed observation point in $\mathbb{R}^d$. Assume that $\omega_x : (0,1] \to \mathbb{R}^+$ is an upper bound on the modulus of continuity of F_x^{-1}. Assume moreover that ω_x is an increasing and continuous function on $(0,1]$ such that $\omega_x(u)/u$ is a non increasing function. Then for any $k \in \{1, \dots, n\}$:*

$$\mathbb{E}\left(|\Delta_{n,\frac{k}{n}}(x)|\right) \leqslant \frac{C}{\sqrt{k}}\left\{\left[F_x^{-1}\left(\frac{k}{n}\right) - F_x^{-1}(0)\right] + \omega_x\left(\frac{\sqrt{k}}{n}\right)\right\} \tag{4.5}$$

$$\leqslant \frac{2C}{\sqrt{k}}\omega_x\left(\frac{k}{n}\right), \tag{4.6}$$

where C is an absolute constant.

This result is derived from deviation bounds we prove in Chazal et al. (2015b). The rate (4.6) can be rewritten as follows:

$$\mathbb{E}\left|\Delta_{n,\frac{k}{n}}(x)\right| \lesssim \frac{n}{k}\frac{1}{\sqrt{n}}\sqrt{\frac{k}{n}}\omega_x\left(\frac{k}{n}\right), \tag{4.7}$$

where the term $\frac{n}{k}$ is the renormalization by the mass proportion $\frac{k}{n}$ in the definition of the DTM, the term $\frac{1}{\sqrt{n}}$ corresponds to a classical parametric rate of convergence and the term $\sqrt{\frac{k}{n}}$ is obtained thanks to a local analysis of the empirical process. More precisely, this last term derives from a sharp control of the variance of a supremum over the uniform empirical process. The term $\omega_x\left(\frac{k}{n}\right)$ corresponds to the statistical complexity of the problem, expressed in term of the regularity of the quantile function F_x^{-1}.

Theorem 10 can be interpreted with either an asymptotic or a non asymptotic point of view. Taking a non asymptotic approach, we consider n as fixed. In the most favorable case where $\tilde{\omega}_x(u) \sim u$, we see in (4.6) that an upper bound of the order of $\frac{1}{n}$ is reached. This is direct consequence of the local analysis we use to control the empirical process in the neighborhood of the origin. Assuming that $\frac{k}{n}$ is very small corresponds to the realistic situation where we use the DTM to clean the support from a small proportion of outliers.

Now, taking an asymptotic approach, Theorem 10 allows us to consider the asymptotic behavior of $\Delta_{n,\frac{k}{n}}(x)$ under all possible regimes, that is for all sequences $(k_n)_{n\in\mathbb{N}}$. For instance, with the classical

approach where k_n is such that $k_n/n = m$ for some fixed value $m \in (0,1)$, we then obtain the parametric rate of convergence $1/\sqrt{n}$. This regime is studied with more details in the following of the chapter.

Another key fact about Theorem 10 is that the upper bound (4.5) depends on the regularity of F_x^{-1} through the function

$$\Psi_x : m \to \frac{\omega_x(m)}{\sqrt{m}}.$$

Moreover, if $\omega(0^+) = 0$, we see that the upper bound (4.5) depends on the regularity of F_x^{-1} only at 0 for n large enough. For instance, if k_n is such that $k_n/n = m$ for some fixed value $m \in (0,1)$ such that $F_x^{-1}(m) > F_x^{-1}(0)$, coming back to (4.5), we find that for n large enough:

$$\omega_x\left(\frac{\sqrt{k_n}}{n}\right) = \omega_x\left(\frac{1}{\sqrt{n}}m\right) < F_x^{-1}(m) - F_x^{-1}(0).$$

In this context, the right hand term of Inequality (4.5) is of the order of $\frac{\widetilde{\Psi}_x(\frac{k_n}{n})}{\sqrt{k_n}}$ where

$$\widetilde{\Psi}_x(m) := \frac{F_x^{-1}(m) - F_x^{-1}(0)}{\sqrt{m}} \qquad \text{for any } m \in (0,1).$$

This remark is confirmed by several numeral experiments we propose in Chazal et al. (2015b).

To complete the results of Theorem 10, we give below a partial lower bound. Let ω be a continuous and increasing function on $[0,1]$ and let $x \in \mathbb{R}^d$. We introduce that class of probability measures:

$$\mathcal{P}_\omega := \left\{ P \text{ is a probability measure on } \mathbb{R}^d \text{ such that } \omega(u) \geqslant \tilde{\omega}_x(u) \text{ for any } u \in (0,1] \right\}.$$

In the previous definition, the function $\tilde{\omega}$ is as before the modulus of continuity of the quantile function of the distribution of the push-forward measure of P by the function $y \mapsto \|y - x\|^r$.

Proposition 2. [Chazal et al. 2015b] *Assume that there exists $P \in \mathcal{P}_\omega$, $c > 0$ and $\bar{u} \in (0,1)$, such that*

$$c\left[F_x^{-1}(u) - F_x^{-1}(0)\right] \geqslant \omega(u) \quad \text{for any } u \in (0, \bar{u}]. \tag{4.8}$$

Then, there exits a constant C which only depends on c, such that for any $k \leqslant \bar{u}n$.

$$\begin{aligned} \sup_{P \in \mathcal{P}_\omega} \mathbb{E}\left(|\Delta_{n,\frac{k}{n},r}(x)|\right) &\geqslant \inf_{\hat{d}_n(x)} \sup_{P \in \mathcal{P}_\omega} \mathbb{E}\left(\left|\hat{d}_n^r(x) - d_{P,m,r}^r(x)\right|\right) \\ &\geqslant C\frac{n}{k}\frac{1}{n}\omega\left(\frac{k-1}{n}\right), \end{aligned}$$

where the infimum is than over all the estimator $\hat{d}_n(x)$ of $d_{P,m,r}(x)$ defined from a sample $X_1, \ldots, X_n$ of distribution P.

This lower bound matches with the upper bound of Theorem 10 when k is very small since it is of the order of $\omega\left(\frac{k}{n}\right)$.

To finish this section, we note that the pointwise convergence rates of Theorem 10 can be easily extended to the sup norm metric over a compact domain $\mathcal{D}$ of $\mathbb{R}^d$: in Chazal et al. (2015b) we obtain for this problem the same rate of convergence up to a $\log n$ factor.

4.2.2 Local analysis of the DTEM in the unbounded case

When the support of P is not bounded, the quantile function F_x^{-1} tends to infinity at 1, the modulus of continuity of F_x^{-1} is not finite and Theorem 10 can not be applied. We now propose a second result under weaker assumptions on the regularity of F_x^{-1}. It shows that under a weak moment assumption, the rate of convergence is the same as for the bounded case, up to a term decreasing exponentially fast to zero.

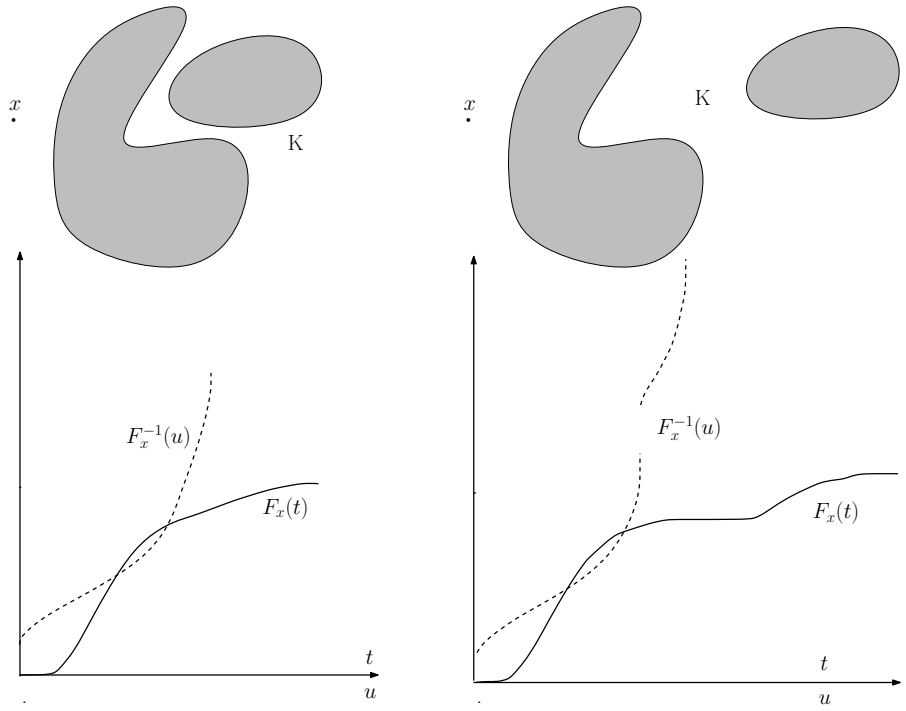


Figure 4.2: Left: one situation where the support of P is not a connected set whereas the support of dF_x is (for $r = 1$). The quantile function F_x^{-1} is continuous. Right: one situation where the support of dF_x is is not a connected set. The quantile function F_x^{-1} is not continuous.

Theorem 11. [Chazal et al. 2015b] *Assume that P has a moment of order r. Let $\bar{m} \in (0,1)$ and some observation point $x \in \mathbb{R}^d$. Assume that $\omega_{x,\bar{m}}$ is an upper bound of the modulus of continuity of F_x^{-1} on $(0,\bar{m}]$: for any $u < u'$ in $[0,\bar{m}]$,*

$$|F_x^{-1}(u) - F_x^{-1}(u')| \leqslant \omega_{x,\bar{m}}(|u - u'|). \tag{4.9}$$

Assume that $\omega_{x,\bar{m}}$ is an increasing and continuous function on $[0,\bar{m}]$ such that $\omega_{x,\bar{m}}(u)/u$ is a non increasing function. Then

$$\mathbb{E}\left|\Delta_{n,\frac{k}{n}}(x)\right| \leqslant \frac{C}{\sqrt{n}}\left[\frac{k}{n}\right]^{-1/2}\left\{\left[F_x^{-1}\left(\frac{k}{n}\right) - F_x^{-1}(0)\right] + \omega_{x,\bar{m}}\left(\frac{\sqrt{k}}{n}\right)\right\} + C_{x,r,\bar{m}}\sqrt{k}\exp\left[-\frac{n^2}{4k}\left(\bar{m} - \frac{k}{n}\right)^2\right]$$

where C is an absolute constant and $C_{x,r,\bar{m}}$ only depends on the quantity $\mathbb{E}\|X - x\|^r$ and on $\bar{m}$.

As for the bounded case, if $\omega(0^+) = 0$ and if $F_x^{-1}(m) > F_x^{-1}(0)$, then the rate of convergence is still of the order of $\frac{\tilde{\Psi}_x(m)}{\sqrt{n}}$. Note that this result is interesting even when the measure P is supported on a compact set. Indeed, assume that the quantile function F_x^{-1} is not continuous, then $\tilde{\omega}_x^{-1}(0) > 0$. However, if F_x^{-1} is smooth in the neighborhood of zero, for $\bar{m}$ small enough the assumption (4.9) may be satisfied with a function $\omega_{x,\bar{m}}$ which can be very small in the neighborhood of zero. Theorem 11 may provide better bounds in this context than those given by Theorem 10.

4.2.3 About the geometric information carried by the quantile function $\mathbf{F_x^{-1}}$

The upper bounds we obtain directly depend on the regularity of F_x^{-1}. We now give some insights about how the geometry of the support of the sampling measure in $\mathbb{R}^d$ impacts the quantile function F_x^{-1}. These remarks are adapted from Appendix 7 in Bobkov and Ledoux (2014).

First, it can be checked that, given the sampling measure P in $\mathbb{R}^d$ and an observation point $x \in \mathbb{R}^d$, the modulus of continuity of the quantile function F_x^{-1} satisfies $\tilde{\omega}_x(u) < \infty$ for any $u \leqslant 1$ if and only if P is compactly supported.

Next, the fact that $\tilde{\omega}_x(0^+) = 0$ is directly related to the connexity of the support of the distribution dF_x. While discontinuity of the distribution function corresponds to atoms, discontinuity points of the quantile function corresponds to area with empty mass in $\mathbb{R}^d$ (see the right picture of Figure 4.2). Indeed, the fact that $\tilde{\omega}_x(0^+) = 0$ is equivalent to assuming that the support of dF_x is a closed interval in $\mathbb{R}^+$ in (see for instance Proposition A.7 in Bobkov and Ledoux, 2014).

In the most favorable situations where the support of P is a connected set, then $\tilde{\omega}_x(0^+) = 0$ and the faster $\tilde{\omega}_x$ tends to 0 at 0, the better the rate we obtain. However, for some point $x \in \mathbb{R}^d$, it is also possible for the support of dF_x to be an interval even when the support of P is not a connected set of $\mathbb{R}^d$ (see the left picture of Figure 4.2). In the other case, when the support of dF_x is not a connected set, the term $\tilde{\omega}_x(0)$ roughly corresponds to the maximum distance between two consecutive intervals of the support of dF_x (see the right picture of Figure 4.2). Our results can still be applied in these situations but the upper bounds we obtain in this case are larger because $\omega_x(\frac{k}{n})$ can not be smaller than $\tilde{\omega}_x(0)$.

Finally, if P be a probability measure on $\mathbb{R}^d$ which is (a, b) standard on its support $\mathbb{X}$ (see Section 3.2), and if $\mathbb{X}$ is a connected set of $\mathbb{R}^d$, then, for any $h \in (0, 1)$ we have $\tilde{\omega}_x(h) \leqslant r \left(\frac{h}{a}\right)^{1/b} \mathrm{d_H}\left(\{x\}, \mathbb{X}\right)^{r-1}$.

4.3 Limiting distribution and bootstrap for the DTM

In this section, we continue the study of the convergence of the DTEM by considering the limiting distribution of the DTM and we show that several bootstrap methods can be applied about the DTM. We start with the pointwise limit in distribution of the DTEM.

Theorem 12. [Chazal et al. 2014a] *Let P be some distribution in $\mathbb{R}^d$. For some fixed x, assume that F_x is differentiable at $F_x^{-1}(m)$, for $m \in (0, 1)$, with positive derivative $F_x'(F_x^{-1}(m))$. Then $\sqrt{n}\Delta_{n,m}$ converges in distribution to $N(0, \sigma_x^2)$, where*

$$\sigma_x^2 = \frac{1}{m^2} \int_0^{F_x^{-1}(m)} \int_0^{F_x^{-1}(m)} [F_x(s \wedge t) - F_x(s)F_x(t)]\, ds\, dt.$$

We also give the functional limit of the DTEM, on a compact domain $\mathcal{D} \subset \mathbb{R}^d$. We say that $\omega_{\mathcal{D}} : (0, 1) \to \mathbb{R}^+$ is an *uniform modulus of continuity* for the family of quantiles functions $(F_x^{-1})_{\mathcal{D}}$ if, for any $u \in (0, 1)$ and any $x \in \mathcal{D}$,

$$\sup_{(m,m') \in (0,1)^2\,,\, |m'-m|<u} |F_x^{-1}(m') - F_x^{-1}(m)| \leqslant \omega_{\mathcal{D}}(u),$$

Here we also assume that

$$\lim_{u \to 0} \omega_{\mathcal{D}}(u) = \omega_{\mathcal{D}}(0) = 0. \tag{4.10}$$

Theorem 13. [Chazal et al. 2014a] *Let P be a measure on $\mathbb{R}^d$ with compact support. Let $\mathcal{D}$ be a compact domain on $\mathbb{R}^d$ and $m \in (0, 1)$. Assume that there exists a uniform modulus of continuity $\omega_{\mathcal{D}}$ for the family $(F_x^{-1})_{\mathcal{D}}$ satisfying (4.10). Then $\sqrt{n}\Delta_{n,m}$ converges in distribution to $\mathbb{B}$ on $\mathcal{D}$, where $\mathbb{B}$ is a centered Gaussian process with covariance kernel*

$$\kappa(x, y) = \frac{1}{m^2} \int_0^{F_x^{-1}(m)} \int_0^{F_y^{-1}(m)} \left(\mathbb{P}\left[B(x, \sqrt{t}) \cap B(y, \sqrt{s}) \right] - F_x(t)F_y(s) \right) ds\, dt.$$

4.3.1 Hadamard differentiability and bootstrap for the DTM

In this section, we use the bootstrap to get a confidence band for the DTM. For some fixed $m \in (0, 1)$, define c_α by

$$\mathbb{P}\left(\sqrt{n}||\Delta_{n,m}||_\infty > c_\alpha\right) = \alpha.$$

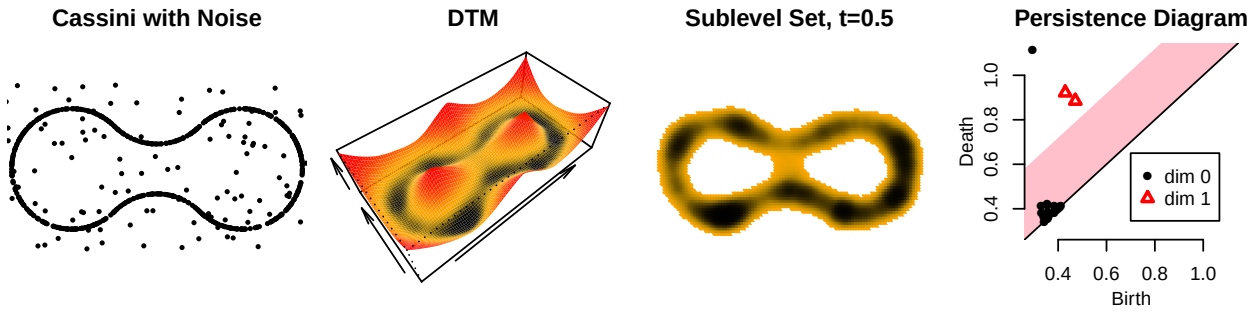


Figure 4.3: The left plot shows a sample from the Cassini curve together with a few outliers. The second plot is the empirical DTM. The third plot is one sub-level set of the DTM. The last plot is the persistence diagram. Points not in the shaded band are significant features. Thus, this method detects one significant connected component and two significant loops.

Let $X_1^*, \ldots, X_n^*$ be a sample from P_n, and let P_n^* be the corresponding (bootstrap) empirical measure. Let us introduce the bootstrap version of $\Delta_{n,m}$: for any $x \in \mathbb{R}^d$,

$$\Delta_{n,m}^*(x) = \Delta_{n,m,r}^*(x) := d_{P_n^*,m}^r(x) - d_{P_n,m}^r(x).$$

The bootstrap estimate $\hat{c}_\alpha$ is defined by

$$\mathbb{P}\left(\sqrt{n}||\Delta_{n,m}^*||_\infty > \hat{c}_\alpha \,|\, X_1, \ldots, X_n\right) = \alpha. \tag{4.11}$$

As usual, $\hat{c}_\alpha$ can be approximated by Monte Carlo. Below we show that this bootstrap is valid.

Theorem 14. [Chazal et al. 2014a] *Let P be a measure on $\mathbb{R}^d$ with compact support $\mathbb{X}$, $m \in (0,1)$ be fixed and $\mathcal{D}$ be a compact domain in $\mathbb{R}^d$. Assume that F_x is differentiable at $F_x^{-1}(m)$ for all $x \in \mathcal{D}$ and that there exist a constant $C > 0$ such that for all small $\eta \in \mathbb{R}$,*

$$\sup_{x \in \mathcal{D}} \left|F_x\left(F_x^{-1}(m)\right) - F_x(F_x^{-1}(m) + \eta)\right| < \epsilon \quad \textit{implies} \quad |\eta| < C\epsilon, \tag{4.12}$$

for all $x \in \mathcal{D}$. Then, $\sup_{x \in \mathcal{D}} \sqrt{n}\left|\Delta_{n,m}^*(x)\right|$ converges in distribution to $\sup_{x \in \mathcal{D}}\left|\frac{1}{m}\int_0^{F_x^{-1}(m)} \mathbb{B}_x(u)du\right|$ conditionally given $X_1, X_2, \ldots,$ in probability.

We establish the above result in Chazal et al. (2015b) using the functional delta method, which entails showing that the distance to measure function is Hadamard differentiable at P. In fact, the proof further shows that the process $x \in \mathcal{D} \mapsto \sqrt{n}\Delta_{n,m}^*$ converges weakly to the Gaussian process $x \in \mathcal{D} \mapsto -\frac{1}{m}\int_0^{F_x^{-1}(m)} \mathbb{B}_x(u)du$. This result is consistent with the result established in Theorem 13, but in order to establish Hadamard differentiability, we use a slightly different assumption. Theorem 13 is proved by assuming an uniform modulus of continuity on the quantile functions F_x^{-1} whereas in Theorem 14 roughly assumes an uniform lower bound on the derivatives of F_x. These two assumptions are consistent: they both say that F_x^{-1} is well behaved in a neighborhood of m for all x. However, (4.12) is stronger.

4.3.2 Bootstrap and significance of topological features

One natural application of DTM in topological data analysis is studying the persistent homology of the sub-levels of the DTM instead of the the sub-levels of the support. Following the ideas of Fasy et al. (2014), we can use the bootstrap to test the significance of a topological feature in the persistence diagram of the sub levels of the DTM. We present two possible methods.

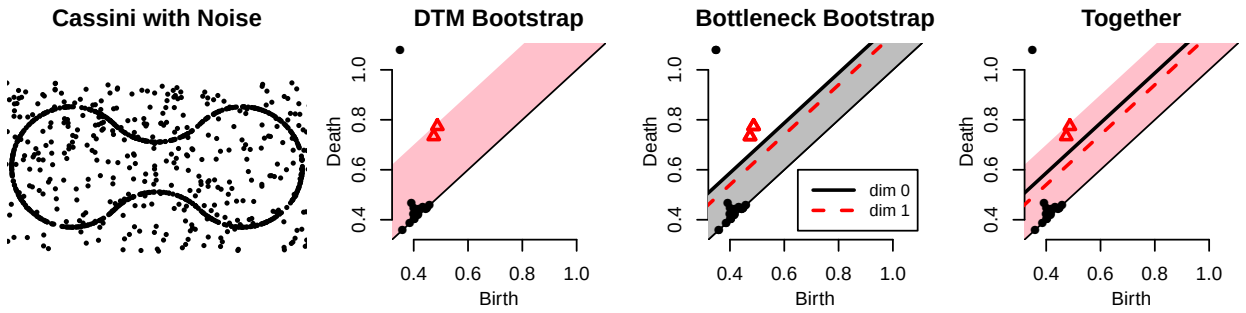


Figure 4.4: The left plot shows a sample from the Cassini curve together with a few outliers. The second plot shows the DTM persistence diagram with a 95% confidence band constructed using the DTM bootstrap. The third plot shows the same persistence diagram with two 95% confidence bands constructed using the bottleneck bootstrap with zero-dimensional features and one-dimensional features. The fourth plot shows the two confidence bands at the same time.

Bootstrapping the DTM. Given a feature with birth and death time (u, v), we will say that the feature is significant if $|v - u| > 2\hat{c}_\alpha/\sqrt{n}$ where $\hat{c}_\alpha$ is defined by (4.11). We now explain why this first method makes sense. Let Dgm be the persistence diagram of the sub-levels of $d_{P,m}$ and the let $\widehat{\text{Dgm}}$ be the persistence diagram of the sub-levels of $d_{P_n,m}$. We introduce the subset of persistence diagrams

$$\mathcal{C}_n = \left\{ E \in \mathcal{D}\text{iag} : \ \mathrm{d_b}(\widehat{\text{Dgm}}, E) \leqslant \frac{\hat{c}_\alpha}{\sqrt{n}} \right\},$$

where $\mathcal{D}$iag is the set of all the persistence diagrams. Then, according to Theorem 14,

$$\mathbb{P}(\text{Dgm} \in \mathcal{C}_n) = \mathbb{P}\left(\mathrm{d_b}(\text{Dgm}, \widehat{\text{Dgm}}) \leqslant \frac{\hat{c}_\alpha}{\sqrt{n}} \right) \geqslant \mathbb{P}\left(\sqrt{n} \|\Delta_{P,m}\|_\infty \leqslant \hat{c}_\alpha \right)$$

where the inequality derives from a stability result due to Buchet et al. (2015b). Now $|v-u| > 2\hat{c}_\alpha/\sqrt{n}$ if and only if the feature cannot be matched to the diagonal for any diagram in $\mathcal{C}$. We can visualize the significant features by putting a band of size $2c_\alpha/\sqrt{n}$ around the diagonal of $\widehat{\text{Dgm}}$, see Fig. 4.3.

The Bottleneck Bootstrap. More precise inferences can be obtained by directly bootstrapping the persistence diagram. Define $\hat{t}_\alpha$ by

$$\mathbb{P}\left(\sqrt{n}\, \mathrm{d_b}(\widehat{\text{Dgm}}^*, \widehat{\text{Dgm}}) > \hat{t}_\alpha \,|\, X_1, \ldots, X_n \right) = \alpha.$$

The quantile $\hat{t}_\alpha$ can be estimated by Monte Carlo. We then use a band of size $2\hat{t}_\alpha$ on the diagram Dgm, see Fig. 4.4. The reason why the bottleneck bootstrap can lead to more precise inferences than the bootstrap from the previous paragrah is that this last uses the fact that $\mathrm{d_b}(\widehat{\text{Dgm}}, \text{Dgm}) \leqslant \|\Delta_{P,m}\|_\infty$ and finds an upper bound for $\|\Delta_{P,m}\|_\infty$. But in many cases the inequality is not sharp so the confidence set can be very conservative.

In Chazal et al. (2014a), we also show similar results for a distance function introduced by Phillips et al. (2014).

4.4 Denoising the DTM via Wasserstein deconvolution

In many situations the observed data is contaminated by an additive noise. In this section it is assumed that we observe n i.i.d. random vectors $(Y_i = (Y_{i,1}, \ldots, Y_{i,d})^t)_{1 \leqslant i \leqslant n}$ of distribution P, in the model

$$Y_i = X_i + \varepsilon_i, \tag{4.13}$$

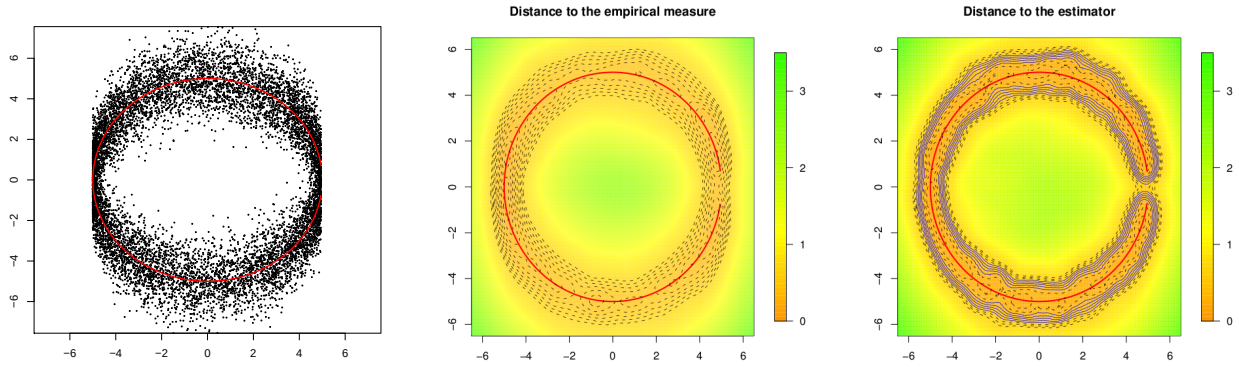


Figure 4.5: Left: circle with hole in red and 10000 points sampled on it with an unidirectional Laplace measurement error. Middle : distance to the empirical measure $d_{P_n,m}$ with $m = 0.01$. Right: distance to the deconvolved measure. The sub-levels of the distance function which have the correct topology are drawn in blue, the other sub-levels are the black dashed lines. A large band of sub-levels of the distance to the denoised measure have the correct topology whereas it is not the case for the DTEM.

where X_i is random vector distributed according to an unknown distribution μ in $\mathbb{R}^d$ and ε_i is the random vector of the noise. In this framework, we would like to infer $d_{\mu,m}$, the distance to the underlying measure μ.

Ideally, we would like to "directly" deconvolve the DTM. However, the DTM being a non linear operator, this direct strategy seems difficult to achieve. In Caillerie et al. (2011), we propose the alternative strategy which consists in first deconvolving the measure and then plugging it in the DTM. This approach is illustrated by Figure 4.5. According to the Wasserstein stability of the DTM, see Inequality (4.2), the convergence of the distance to the deconvolved measure can be derived from the convergence of deconvolved measure under Wasserstein metrics. This last problem is the main subject of this section.

Besides the geometric applications we have in mind, studying the properties of probability estimators for the Wasserstein metric is also interesting in itself. Firstly, contrary to the $\mathbb{L}_r$-distances between probability densities (except for $r = 1$, which coincides with the total variation distance), the distances W_r are true distances between probability distributions. Secondly, many natural estimators $\hat{Q}_n$ of Q are singular with respect to Q (think of the empirical measure in most cases), and consequently the total variation distance between $\hat{Q}_n$ and Q is equal to 2 for any n. This is the case of our deconvolution estimator, if the support $\mathbb{X}$ is a submanifold in $\mathbb{R}^d$ with dimension strictly less than d. Wasserstein metrics appear as natural distances to evaluate the performance of such estimators.

4.4.1 Deconvolution of a measure and Wasserstein metric

We first give more precisions on the framework of this work. We assume that the random vectors $X_i = (X_{i,1}, \ldots, X_{i,d})^t$ in the model (4.13) are i.i.d and distributed according to μ supported on an unknown compact subset $\mathbb{X}$ of $\mathbb{R}^d$. The random vectors $\varepsilon_i = (\varepsilon_{i,1}, \ldots \varepsilon_{i,d})^t$'s are also i.i.d. random and distributed according to a probability measure μ_ε. The observations are thus drawn according to the convolution model $P = \mu \star \mu_\varepsilon$. For the applications we have in mind, μ is typically supported by a submanifold of $\mathbb{R}^d$. Consequently, we shall not assume that μ has a density with respect to the Lebesgue measure on $\mathbb{R}^d$.

In this section, we present the case where the coordinates of the error vectors are independent. In Caillerie et al. (2011); Dedecker and Michel (2013); Dedecker et al. (2015), we have also considered the more general case where the noise vector is a linear transform of errors with independent coordinates.

We now introduce the deconvolution estimator. For $r \in [1, \infty[$ let $\lceil r \rceil$ be the smallest integer greater than r. We first define a kernel k whose Fourier transform is smooth enough and compactly supported over $[-1, 1]$. Such kernels can be defined by considering powers of the sinc function. More precisely,

let

$$k(x) = c_r \left\{ \frac{(2\lceil r/2\rceil + 2)\sin\frac{x}{2\lceil r/2\rceil+2}}{x} \right\}^{2\lceil r/2\rceil+2},$$

where c_r is such that $\int k(x)dx = 1$. For any $j \in \{1, \cdots, d\}$ and any $h_j > 0$, let

$$\tilde{k}_{j,h_j}(x) = \frac{1}{2\pi}\int e^{iux}\frac{k^*(u)}{{\mu_{\varepsilon,j}}^*(u/h_j)}du \tag{4.14}$$

where k^* and $\mu^*_{\varepsilon,j}$ are the Fourier transform of k and $\mu_{\varepsilon,j}$. A preliminary estimator $\hat{f}_n$ is given by

$$\hat{f}_n(x_1,\ldots,x_d) = \frac{1}{n}\sum_{i=1}^{n}\prod_{j=1\ldots d}\frac{1}{h_j}\tilde{k}_{j,h_j}\Big(\frac{x_j - Y_{i,j}}{h_j}\Big). \tag{4.15}$$

The estimator (4.15) is the multivariate version of the standard deconvolution kernel density estimator which was first introduced in Carroll and Hall (1988). The estimator $\hat{f}_n$ is not necessarily a density, since it has no reason to be non negative. Since our estimator has to be a probability measure, we define

$$\hat{g}_n(x) = \alpha_n \hat{f}_n^+(x), \quad \text{where} \quad \alpha_n = \frac{1}{\int_{\mathbb{R}^d}\hat{f}_n^+(x)dx} \quad \text{and} \quad \hat{f}_n^+ = \max\{0, \hat{f}_n\}.$$

The estimator $\hat{\mu}_n$ of μ is then the probability measure with density $\hat{g}_n$.

4.4.2 Rates of convergence

Minimax rates of convergence for deconvolving an univariate density have been deeply studied, see for instance Fan (1991); Butucea and Tsybakov (2008a,b); Meister (2009). The multivariate problem has also been investigated by Tang (1994) and by Comte and Lacour (2011). All these contributions concern pointwise convergences or $\mathbb{L}^2$ convergences whereas our contributions dealt with rates of convergence for the Wasserstein metrics.

From a bias-variance decomposition of $\hat{f}_n$, we obtain a general upper bound for the estimation of μ for the W_r risk. As usual in deconvolution problems, the rates of convergence depends on the derivatives of the functions $t_j := 1/\mu^*_{\varepsilon,j}$.

Proposition 3. [Dedecker and Michel 2013] *Let $(h_1,\ldots,h_d) \in [0,1]^d$. The following upper bound holds*

$$\mathbb{E}_{(\mu\star\mu_\varepsilon)^{\otimes n}}(W_r^r(\hat{\mu}_n,\mu)) \leqslant (2d)^{r-1}\left(\int|u|^r k(u)du\right)(h_1^r+\cdots+h_d^r) + \frac{L}{\sqrt{n}}\left(\prod_{j=1}^{d} I_j(h_j) + \sum_{\ell=1}^{d} J_\ell(h_\ell)\Big(\prod_{j=1,j\neq\ell}^{d} I_j(h_j)\Big)\right)$$

where L is some positive constant L and

$$\begin{aligned} I_j(h) &\leqslant \sqrt{\int_{-1/h}^{1/h}(t_j(u))^2 + (t_j'(u))^2 du}, \\ J_j(h) &\leqslant \sqrt{\int_{-1/h}^{1/h}\Big(t_j(u))^2 + (t_j^{([r]+1)}(u)\Big)^2 du} + \sum_{k=1}^{[r]} h^{[r]+1-k}\sqrt{\int_{-1/h}^{1/h}\Big(t_j^{(k)}(u)\Big)^2 du}. \end{aligned}$$

For some $M > 0$ and $r \geqslant 1$, we consider the deconvolution of measures μ on $\mathbb{R}^d$ such that

$$\sup_{1\leqslant j\leqslant d}\mathbb{E}_\mu\Big((1+|(X_1)_j|^{2r+2})\prod_{1\leqslant\ell\leqslant d,\,\ell\neq j}(1+|(X_1)_\ell|^2)\Big) \leqslant M.$$

We denote by $\mathcal{D}(M,r)$ the set of all the measures μ satisfying this condition.

Super smooth distributions

In the case where ε_1 is a non degenerate Gaussian random vector, the minimax rate of convergence of the sets $\mathcal{D}(M,r)$ is of the order of $(\log n)^{r/2}$:

Theorem 15. [Dedecker and Michel 2013] *Assume that we observe $Y_1,\dots,Y_n$ in the convolution model (4.13), where ε_1 is a non degenerate Gaussian random vector. Let $M>0$ and $p\geqslant 1$. Then*

$$\sup_{n\geqslant 1}\sup_{\mu\in\mathcal{D}(M,r)}(\log n)^{r/2}\ \mathbb{E}_{(\mu\star\mu_\varepsilon)^{\otimes n}}(W_r^r(\hat{\mu}_n,\mu))\leqslant K$$

for some positive constant K. Moreover, there exists a constant $C>0$ such that for any estimator $\tilde{\mu}_n$ of the measure μ:

$$\liminf_{n\to\infty}(\log n)^{r/2}\sup_{\mu\in\mathcal{D}(M,r)}\mathbb{E}_{(\mu\star\mu_\varepsilon)^{\otimes n}}(W_r^r(\tilde{\mu}_n,\mu))\geqslant C.$$

The upper bound derives from Proposition 3 in the particular case of Gaussian noise. The lower bounds (in any dimension) can be deduced from lower bounds for the deconvolution of the cumulative distribution function in dimension one. We have proved a more general result for all supersmooth distributions in Dedecker and Michel (2013).

Ordinary smooth distributions

Contrary to the supersmooth case, for ordinary smooth distributions the rate of convergence depends on the dimension. Proving minimax rates of convergence is much more challenging in this case. In the multivariate convolution model (4.13) where the marginal distributions μ_{ε_j}'s all have a Laplace distribution, we deduce from Proposition 3 that

$$W_r^r(\hat{\mu}_n,\mu)\lesssim n^{-\frac{r}{2r+5d}}. \tag{4.16}$$

Proving comparable lower bounds in this context is an open problem.

In the unidimensional case however, in our paper Dedecker et al. (2015) we obtain a better upper bound then (4.16) for an alternative estimator based on the estimation of the cumulative distribution function F of μ. This estimator $\tilde{\mu}_n$ is built in two steps:

1. **A preliminary estimator of F.** We define a preliminary estimator $\hat{F}_n$ of F:

$$\hat{F}_n(t)=\frac{1}{nh}\int_{-\infty}^{t}\sum_{k=1}^{n}\tilde{k}_h\left(\frac{u-Y_k}{h}\right)du$$

   where $\tilde{k}_h$ corresponds to $\tilde{k}_h$ in (4.14) in this unidimensional framework. Note that $\hat{F}_n$ is not a cumulative distribution function since it is not necessarily non-decreasing.

2. **Isotone approximation.** We need to define an estimator $\tilde{F}_n$ of F which is a cumulative distribution function. Let $\tilde{F}_n$ be such that, for every distribution function G,

$$\int|x|^{r-1}|\hat{F}_n-\tilde{F}_n|(x)dx\leqslant\int|x|^{r-1}|\hat{F}_n-G|(x)dx+n^{-1/2}$$

   The estimator $\tilde{\mu}_n$ is then defined as the probability measure with distribution function $\tilde{F}_n$.

Let m_0 denote the least integer strictly greater than $r+\frac{1}{2}$, and m_1 be the least integer strictly greater than $r-\frac{1}{2}$. Let $t_\varepsilon=1/\mu_\varepsilon^*$.

Theorem 16. [Dedecker et al. 2015] *Assume that t_ε is at least m_1 times continuously differentiable. Assume that*

$$\int_0^\infty|x|^{r-1}\sqrt{P(|Y|\geqslant x)}dx<\infty\ \text{ and }\ \sup_{u\in[-2,2]}|t_\varepsilon^{(m_0)}(u)|<\infty.$$

Also assume that there exist $\beta > 0$ and $c > 0$, such that for every $\ell \in \{0, 1, \ldots, m_1\}$ and every $u \in \mathbb{R}$,

$$|t_\varepsilon^{(\ell)}(u)| \leqslant c(1 + |u|)^\beta.$$

Then, taking $h = n^{-\frac{1}{2r+(2\beta-1)_+}}$, there exists a positive constant C such that

$$\mathbb{E}W_r^r(\tilde{\mu}_n, \mu) \leqslant C\psi_n \quad \textit{where} \quad \psi_n = \begin{cases} n^{-\frac{r}{2r+2\beta-1}} & \textit{if } \beta > \frac{1}{2} \\ \sqrt{\frac{\log n}{n}} & \textit{if } \beta = \frac{1}{2} \\ \frac{1}{\sqrt{n}} & \textit{if } \beta < \frac{1}{2} \end{cases}. \tag{4.17}$$

Applying Proposition 3 in the same context of Theorem 16 gives an upper bound of the order of $n^{-r/(2r+2\beta+1)}$ for the risk of the previous estimator $\bar{\mu}_n$, which is worse than (4.17). However, contrary to $\hat{\mu}_n$, the estimator $\tilde{\mu}_n$ is well defined for $d = 1$ only.

We give below a lower bound that partially matches with the upper bounds of Theorem 16. Let $\mathcal{D}_q(M)$ be the set of measures μ on $\mathbb{R}$ such that $\int |x|^q d\mu(x) \leqslant M$.

Theorem 17. [Dedecker et al. 2015] *Let $M > 0$ and $q \geqslant 1$. Assume that t_ε is at least two times continuously differentiable and that there exist $\beta > 0$ and $c > 0$, such that for every $\ell \in \{0, 1, 2\}$ and every $u \in \mathbb{R}$,*

$$|\mu_\varepsilon^{*(\ell)}(u)| \leqslant c(1 + |u|)^{-\beta}.$$

Then, there exists a constant $C > 0$ such that, for any estimator $\hat{\mu}$,

$$\liminf_{n\to\infty} n^{\frac{r}{2\beta+1}} \sup_{\mu\in\mathcal{D}_q(M)} \mathbb{E}W_r^r(\hat{\mu}, \mu) > C.$$

For W_1, this lower bound matches the upper bound given in Theorem 16 for $\beta \geqslant 1/2$. For W_r ($r > 1$), we conjecture that the upper bounds given by Theorem 16 are appropriate under the assumed tail conditions.

4.5 Discussion and directions for future research

Statistical analysis of the DTM and its variants

Our contributions about the convergence of the DTEM is a step toward a complete statistical analysis of robust geometric inference. One first objective for the future is to provide a better lower bound than the one given by Proposition 2. We also would like to be able to control the convergence of the DTEM uniformly on $\mathbb{R}^d$ rather than over compact sets.

One natural application of the DTM is estimating the support of a distribution, when this distribution is contaminated by noise. In some situations, the distribution of the noise can be learned from the data (see below) and then deconvolution methods can be applied. However this strategy is not always possible. One idea in this more general setting would be to choose m in the DTEM to recover the distance to the support, as well as possible, without using deconvolution methods. Our results on $\Delta_{n,m,r}$ are not sufficient to answer this problem (the bias term is not considered) but some numerical experiments presented in Chazal et al. (2015b) show that the term $\mathbb{E}\Delta_{n,m,r}(x)$ does not have a typical monotonic behavior with regard to m and thus classical model selection methods can be hardly applied to this problem. We intend to study this non standard model selection problem in future works.

We also intend to study an approximation of the DTEM which has been proposed by Guibas et al. (2013), the *witnessed k-distance*, for which level sets are easier to compute. In particular, we would like to study the convergence of the witnessed k-distance to the DTM.

Other applications of the DTM

The potential applications of the DTM are not restricted to the fields of topological data analysis and support estimation. The DTM is a general tool for exploring point clouds. We intend to develop methods based on the DTM for anomaly detection, for comparison of point clouds and also for variable selection.

Wasserstein deconvolution

Many problems related to Wasserstein deconvolution remain unsolved. Firstly, the minimax rate of convergence in the ordinary case for $d > 1$ is still an open question. This is not an easy question, note that lower bounds (not in the minimax sense) for the noiseless case have been proved only very recently by Dereich et al. (2013).

Secondly, we intend to adapt the works of Delaigle et al. (2008) for tuning the bandwidth with bootstrap methods. This requires to compute Wasserstein distances, hopefully recent advances have been maid recently on this problem (Mérigot, 2011).

Finally, in our results the noise distribution is always assumed to be known. Of course, this is not a realistic assumption for the applications. We would like to learn the noise distribution from the data, as proposed in Neumann and Hössjer (1997) for the case of the $\mathbb{L}^2$ metric. However, it seems that adapting the results of Neumann and Hössjer (1997) to our context is not an obvious problem.

Part II

Other contributions in the field of Statistics

Chapter 5

Gaussian mixture clustering

Clustering methods consists of discovering clusters among observations. Many cluster analysis methods have been proposed in statistics and learning theory, one of the most popular approach in this field is model-based clustering. Model-based clustering methods define clusters as observations having most likely the same distribution. In this framework, the distribution of each subpopulation is modeled by a parametric density, like a Gaussian one and thus the unknown data density is estimated by a mixture of these distributions (McLachlan and Peel, 2000). The data clustering is deduced thanks to the maximum a posteriori (MAP) rule and the clustering problem reduces to a density estimation problem.

Cluster analysis is more and more concerned with large datasets where observations are described by many variables. This large number of variables could be beneficial but in many situations, the presence of noisy variables can be harmful to detect a reasonable clustering structure. In our work, the variables are partitioned into two categories. The subset $\mathbf{v}^c$ contains the noisy variables, said irrelevant in the sequel. The distribution of such a noisy variable is assumed to be homogeneous and centered around its mean, allowing not to distinguish a possible clustering of the data. The complementary set $\mathbf{v}$ is composed of the clustering variables.

Because of their wide range flexibility, Gaussian mixture densities are widely used to model the unknown distribution of continuous data for clustering analysis (Lindsay, 1995; McLachlan and Peel, 2000). This chapter presents our contributions on the problem of selecting a convenient Gaussian mixture model (GMM) for clustering. The first section presents our results about a non asymptotic model selection method based on a ℓ_0 penalty, they were obtained during my Phd Thesis (Michel, 2008; Maugis and Michel, 2011b,a). In the continuity of this work, we later proved the adaptivity of the method in the univariate case (Maugis-Rabusseau and Michel, 2013).

In all the chapter, we observe a sample $(X_1, \ldots, X_n)$ of i.i.d. random vectors from a distribution with density s in $\mathbb{R}^d$.

5.1 Gaussian mixture selection through ℓ_0 penalization

Let $\mathcal{V}$ be the collection of nonempty subsets of $\{1, \ldots, d\}$. A Gaussian mixture family is characterized by the number of components $K \in \mathbb{N}^*$ and the relevant variable index subset $\mathbf{v} \in \mathcal{V}$ whose cardinal is denoted α. In the sequel, the set of index couples $(K, \mathbf{v})$ is $\mathcal{M} = \mathbb{N}^* \times \mathcal{V}$. Consider the decomposition of a vector $x \in \mathbb{R}^d$ into its restriction on relevant variables $x_{[\mathbf{v}]} = (x_{j_1}, \ldots, x_{j_\alpha})^t$ and its restriction on irrelevant variables $x_{[\mathbf{v}^c]} = (x_{l_1}, \ldots, x_{l_{d-\alpha}})^t$ where $\mathbf{v} = \{j_1, \ldots, j_\alpha\}$ and $\mathbf{v}^c = \{l_1, \ldots, l_{d-\alpha}\} = \{1, \ldots, d\} \backslash \mathbf{v}$. On clustering variables, a Gaussian mixture f is chosen among the following mixture family

$$\mathcal{L}_{(K,\alpha)} = \left\{ \sum_{k=1}^{K} p_k \Phi(.|\mu_k, \Sigma_k); \begin{array}{l} \forall k,\ \mu_k \in [-a, a]^\alpha,\ (\Sigma_1, \ldots, \Sigma_K) \in \mathcal{D}^+_{(K,\alpha)} \\ 0 < p_k < 1, \sum_{k=1}^{K} p_k = 1 \end{array} \right\}$$

where $a \in \mathbb{R}^*_+$ and $\mathcal{D}^+_{(K,\alpha)}$ denotes a family of K-tuples of $\alpha \times \alpha$ symmetric positive definite matrices which corresponds to the chosen Gaussian mixture form. On irrelevant variables, the data density is

modeled by a spherical Gaussian density g belonging to the following family

$$\mathcal{G}_{(\alpha)} = \left\{\Phi(.|0, \omega^2 I_{d-\alpha});\ \omega^2 \in [\lambda_{\mathrm{m}}, \lambda_{\mathrm{M}}]\right\}$$

where $0 < \lambda_{\mathrm{m}} < \lambda_{\mathrm{M}}$. Finally, the family of Gaussian mixture associated to $(K, \mathbf{v}) \in \mathcal{M}$ is defined by

$$\mathcal{S}_{(K,\mathbf{v})} = \left\{x \in \mathbb{R}^d \mapsto f(x_{[\mathbf{v}]})\, g(x_{[\mathbf{v}^c]});\ f \in \mathcal{L}_{(K,\alpha)},\ g \in \mathcal{G}_{(\alpha)}\right\}.$$

The dimension of the model $\mathcal{S}_{(K,\mathbf{v})}$ is denoted $D(K, \mathbf{v})$ and corresponds to the free parameter number of Gaussian mixtures in this model. It only depends on the number K of components, the Gaussian mixture form and the number of clustering variables α. We study various forms of Gaussian mixtures in Maugis and Michel (2011b), based on the eigenvalue decomposition of the variance matrices as in Banfield and Raftery (1993).

The maximum likelihood estimator of the density on $S_{(K,\mathbf{v})}$ is defined by $\hat{s}_{(K,\mathbf{v})} := \underset{t \in S_{(K,\mathbf{v})}}{\operatorname{argmin}} \gamma_n(t)$ for the empirical contrast $\gamma_n(t) = -\frac{1}{n}\sum_{i=1}^n \ln\{t(X_i)\}$. The risk of an estimator $\hat{s}_{(K,\mathbf{v})}$ is defined by

$$\mathcal{R}(\hat{s}_{(K,\mathbf{v})}) = \mathbb{E}[\mathrm{KL}(s, \hat{s}_{(K,\mathbf{v})})],$$

where KL is the Kullback-Leibler divergence. We also use the notation $\mathcal{H}$ for the Hellinger distance in this chapter.

Starting from general results on model selection from Bigé and Massart (Massart, 2007) and by computing the bracketing entropies of these multivariate gaussian mixture models, we prove the following oracle inequalities for various different forms of GMM (among others shapes, we consider diagonal and general GMM):

Theorem 18. [Maugis and Michel 2011b] *For one given collection of GMM:*

1. *If the variables are ordered, there exists two absolute constants κ and C such that, if*

$$\mathrm{pen}(K, \mathbf{v}) \geqslant \kappa \frac{D(K, \mathbf{v})}{n}\left[2\,A + \ln\left(\frac{1}{1 \wedge \frac{D(K,\mathbf{v})}{n} A}\right) + 1\right] \tag{5.1}$$

then the model $(\hat{K}, \hat{\mathbf{v}})$ minimizing $\mathrm{crit}(K, \mathbf{v}) = \gamma_n(\hat{s}_{(K,\mathbf{v})}) + \mathrm{pen}(K, \mathbf{v})$ on $\mathcal{M}$ exists and

$$\mathbb{E}\left[\mathcal{H}^2(s, \hat{s}_{(\hat{K},\hat{\mathbf{v}})})\right] \leqslant C\left\{\inf_{(K,\mathbf{v})\in\mathcal{M}}[\mathrm{KL}(s, \mathcal{S}_{(K,\mathbf{v})}) + \mathrm{pen}(K, \mathbf{v})] + \frac{1}{n}\right\}. \tag{5.2}$$

2. *If the variables are not ordered, there exists two absolute constants κ and C such that, if*

$$\mathrm{pen}(K, \mathbf{v}) \geqslant \kappa \frac{D(K, \mathbf{v})}{n}\left\{2\,A + \ln\left[\frac{1}{1 \wedge \frac{D(K,\mathbf{v})}{n} A}\right] + \frac{1}{2}\ln\left[\frac{8\exp(1)d}{(D(K, \mathbf{v}) - 1) \wedge (2d - 1)}\right]\right\},$$

then the model $(\hat{K}, \hat{\mathbf{v}})$ minimizing $\mathrm{crit}(K, \mathbf{v}) = \gamma_n(\hat{s}_{(K,\mathbf{v})}) + \mathrm{pen}(K, \mathbf{v})$ on $\mathcal{M}$ exists and

$$\mathbb{E}\left[\mathcal{H}^2(s, \hat{s}_{(\hat{K},\hat{\mathbf{v}})})\right] \leqslant C\left\{\inf_{(K,\mathbf{v})\in\mathcal{M}}[\mathrm{KL}(s, \mathcal{S}_{(K,\mathbf{v})}) + \mathrm{pen}(K, \mathbf{v})] + \frac{2}{n}\right\}. \tag{5.3}$$

Moroever, $A = O(\sqrt{\ln d})$ as d tends to infinity.

The constant A is a function of the parameters of the models (box width, lower bound on the covariance spectrum,... depending on the shape we choose). The penalty functions take the model complexity into account through $D(K, \mathbf{v})$ as well as the richness of model family. Indeed in the non-ordered variable case, the number of models with the same dimension is larger, and the associated penalty functions have an additional logarithm term depending on the dimension.

Theorem 18 gives the general form of penalty functions but it does not provide explicit penalties since (5.2) and (5.3) depend on absolute unknown constants and mixture parameters are not bounded

in practice. In Maugis and Michel (2011a), we apply the slope heuristics method introduced by Birgé and Massart (2007) to calibrate these penalties. The slope heuristics is presented in the next chapter. Through this way, we obtain a completely data driven model selection method. An intensive experiment study is proposed in Maugis and Michel (2011a) with applications to real data in genomics and curve clustering. It confirms the efficiency of the clustering method. The slope heuristics method and its implementation is presented in the next Chapter.

Several improvements by other authors have been proposed in the following of our works: Cohen and Le Pennec (2013) generalize our results for all possible GMM shapes and also propose a penalized estimator conditional density estimation. Meynet and Maugis-Rabusseau (2012) introduce a new model selection method on GMM which first consists of using a ℓ_1-regularization method to build a data-driven model subcollection and next select a convenient model with a ℓ_0-penalty.

5.2 Minimax adaptivity in the unidimensional case

In Maugis-Rabusseau and Michel (2013), we study the adaptivity of the penalized estimator defined in Theorem 18 in the unidimensional case. In this context, only the number of components K in the mixture are selected. More precisely, we consider the set of densities of the form

$$S_K = \left\{ x \in \mathbb{R} \mapsto \sum_{k=1}^{K} p_k \psi_\sigma(x-\mu_k); \mu_k \in [-\bar{\mu},\bar{\mu}], p_k \in [0,1], \sum_{k=1}^{K} p_k = 1 \right\}$$

where ψ_σ is the univariate centered Gaussian density with variance σ^2. For $\beta > 0$ let $r = \lfloor \beta \rfloor$ be the largest integer less than β and let $k \in \mathbb{N}$ such that $\beta \in (2k, 2k+2]$. We define $\mathbb{H}(\beta)$ as the set of densities on $\mathbb{R}$ whose logarithm is locally β-Hölder: for all x and y such that $|y-x| \leqslant \gamma$,

$$\left| (\ln f)^{(r)}(x) - (\ln f)^{(r)}(y) \right| \leqslant r!\, L(x) |y-x|^{\beta-r}.$$

where γ is a fixed positive parameter and L is a polynomial function on $\mathbb{R}$. Other tail, moments and monotonicity conditions are also required to define $\mathbb{H}(\beta)$ but we omit them for this presentation.

First, we prove a lower bound on the minimax risk for the estimation of density in $\mathbb{H}(\beta)$: we show that the minimax risk is lower bounded by the rate $n^{-\frac{2\beta}{2\beta+1}}$ on the density sets $\mathbb{H}(\beta)$. For this task, we combine a corollary of a Birgé's Lemma (Birgé, 2005) and the so-called Varshamov-Gilbert's Lemma (see for instance Corollary 2.19 and Lemma 4.7 in Massart, 2007). Next, we prove an approximation result between densities in $\mathbb{H}(\beta)$ and the univariate Gaussian mixtures, in KL divergence by adapting a result of Kruijer et al. (2010). This approximation result allows us to control the bias term in the oracle inequality (5.2) (in the unidimensional framework). By taking σ of the order of $K^{-1}(\ln K)^{3/2}$ and $\bar{\mu}$ of the order of $|\ln \lambda(K)|^{1/2}$ in the definition of the models S_K, we finally obtain that the risk of the penalized estimator is bounded by a rate of the order of $(\ln n)^{\frac{5\beta}{2\beta+1}}\, n^{\frac{-2\beta}{2\beta+1}}$. This shows the minimax adaptivity of $\hat{s}_{\hat{K}}$ on the sets $\mathbb{H}(\beta)$, up to a logarithm term.

5.3 Discussion and directions for future research

As far as we know, our paper Maugis-Rabusseau and Michel (2013) has been the first result on the adaptivity of maximum likelihood estimators for Gaussian mixtures. However, regarding the clustering problem, it is not completely satisfactory because it concerns the estimation of a density instead of considering a risk for the clustering problem. Advances have been proposed more recently in this direction in Azizyan et al. (2013) and in Arias-Castro and Verzelen (2014).

In a work in progress with S. Gaiffas, we propose an alternative method based on a PAC Bayesian approach (Gaiffas and Michel, 2014). More precisely we use a generalized Bayesian posterior with a sparsity inducing prior on K and $\mathbf{v}$. We prove a sparsity oracle inequality which shows that this procedure selects the optimal parameters K and $\mathbf{v}$. The method is implemented using a Metropolis-Hastings algorithm, based on a clustering-oriented greedy proposal.

Chapter 6

Slope Heuristics and the Capushe package

Many algorithms in statistics depend on free parameters which can be difficult to tune in practice. To answer this question, Birgé and Massart (2007) have proposed the slope heuristics method[1]. First introduced in the framework of Gaussian regression with a homoscedastic fixed design, it has then been generalized in the heteroscedastic random-design case (Arlot and Massart, 2009). It has also been validated for least squares density estimation (Lerasle, 2012) and for maximum likelihood estimation density estimation (Saumard, 2010) Its practical validity has been illustrated in many frameworks: change-point detection in a Gaussian least squares framework (Lebarbier, 2005), simultaneous variable selection and clustering in a Gaussian mixture models setting (Chapter 5), Gaussian Markov random field framework (Verzelen, 2010) and computational geometry (Chapter 2) to cite a few.

This chapter presents our contribution to the slope heuristics problem: an efficient implementation of this calibration method with the Matlab package CAPUSCHE (Baudry et al., 2012) and with the R package CAPUSCHE (Brault et al., 2011). Note that another objective of our paper Baudry et al. (2012) was to introduce the slope heuristics to a large audience of applied statisticians.

6.1 Contrast minimization and slope heuristics

Let us first briefly recall the general framework of estimation by contrast minimization. The previous chapter provides an illustration of this method is the context of Gaussian mixtures. Let $\mathbf{X} = (X_1, \ldots, X_n)$, $X_i \in \mathbb{R}^d$, be an i.i.d sample from an unknown probability distribution. The quantity of interest, denoted as s (for instance the density of the distribution), is related to the unknown sample distribution and belongs to a set $\mathcal{S}$. The method is based on the existence of a *contrast* function $\gamma : \mathcal{S} \times \mathbb{R}^d \to \mathbb{R}$ fulfilling the fundamental property that

$$s = \underset{t \in \mathcal{S}}{\operatorname{argmin}} \, \mathbb{E}_X \left[\gamma(t, X) \right],$$

where the expectation is taken with respect to X distributed as the sample. The associated *loss function*, which enables us to evaluate each element of $\mathcal{S}$, is defined by:

$$\forall t \in \mathcal{S}, \; l(s,t) = \mathbb{E}_X \left[\gamma(t, X) \right] - \mathbb{E}_X \left[\gamma(s, X) \right].$$

The empirical contrast is defined by $\forall t \in \mathcal{S}, \; \gamma_n(t) = \frac{1}{n} \sum_{i=1}^{n} \gamma(t, X_i)$. Let S be a model, namely a subset of $\mathcal{S}$. A minimizer of the empirical contrast over the model S is then considered and denoted as $\hat{s}$. It is expected that $\hat{s}$ is a sensible estimator of s since, under reasonable conditions, $\gamma_n(t)$ converges to $\mathbb{E}[\gamma(t, X)]$. The quality of such an estimator can be measured by its *risk* $\mathcal{R}(\hat{s}) = \mathbb{E}_{\mathbf{X}} \left[l(s, \hat{s}) \right]$.

In practice, several estimators can be proposed to estimate s. Formally, a countable collection of models $(S_m)_{m \in \mathcal{M}}$ with the corresponding estimators collection $(\hat{s}_m)_{m \in \mathcal{M}}$ of s is now considered. Let $S_{\hat{m}}$ be the model selected by a given model selection procedure. The selected estimator is then $\hat{s}_{\hat{m}}$, where both $\hat{s}_m$ (for any m) and $\hat{m}$ are built from the same sample $\mathbf{X}$. From a non asymptotic point of view, the ideal model S_{m*} for a given n and a given dataset is such that

$$m^* \in \underset{m \in \mathcal{M}}{\operatorname{argmin}} \, l(s, \hat{s}_m). \tag{6.1}$$

[1] the slope heuristics takes its name from a slope estimation.

The aim is to build a model selection procedure such that the selected model $S_{\hat{m}}$ is *optimal* in the sense that it fulfills an oracle inequality:

$$\mathbb{E}_{\mathbf{X}}\left[l(s,\hat{s}_{\hat{m}})\right] \leqslant A_n \inf_{m\in\mathcal{M}} \mathbb{E}_{\mathbf{X}}\left[l(s,\hat{s}_m)\right] + \eta_n.$$

with η_n a small remainder term.

Penalization consists of defining a proper penalty function $\text{pen} : \mathcal{M} \longrightarrow \mathbb{R}^+$ and of selecting $\hat{m}$ minimizing the associated penalized criterion

$$\forall m \in \mathcal{M},\ \text{crit}(m) = \gamma_n(\hat{s}_m) + \text{pen}(m). \tag{6.2}$$

Choosing the penalty is tricky but obviously crucial. Some well-known penalized criteria with fixed penalties such as AIC (Akaike, 1973) or BIC (Schwarz, 1978) have been widely studied (Burnham and Anderson, 2002). The use of these penalties is mainly motivated by asymptotic arguments that may be wrong in a non asymptotic context. More recent works based on concentration inequalities have led to optimal penalties which are known up to a multiplicative constant κ. In this framework, the penalty shape is then denoted as $\text{pen}_{\text{shape}}(\cdot)$ and an unknown constant κ_{opt} exists such that

$$\text{pen}_{\text{opt}} : m \in \mathcal{M} \mapsto \kappa_{\text{opt}}\text{pen}_{\text{shape}}(m) \tag{6.3}$$

is an optimal penalty. Two different kinds of results usually lead to such a penalty shape:

- **Deterministic penalty shapes.** Specific deterministic functions $m \mapsto \text{pen}_{\text{shape}}(m)$ can be used to define an optimal penalty (see Massart, 2007, for some examples of such penalties). For instance, in a general maximum likelihood framework, Theorem 7.11 in Massart (2007) provides a solution to choose a penalty shape and insures the existence of a constant κ_{opt} such that $\text{pen}_{\text{opt}}(\cdot) = \kappa_{\text{opt}}\,\text{pen}_{\text{shape}}(\cdot)$ follows an oracle inequality. The value of κ_{opt} which can be derived from the theory is much too pessimistic and a reasonable value has to be guessed from the data.

- **Resampling penalty shapes.** In a regression framework, Arlot (2009) uses resampling to design the penalty corresponding to each model and derives non asymptotic results for the corresponding procedures. These penalties actually have to be calibrated by a multiplicative constant. Lerasle (2012) provides analogous results in a density estimation framework.

Birgé and Massart (2007) proposes a practical method based on theoretical results for defining efficient penalty functions from the data. We give in Baudry et al. (2012) a non technical presentation of the ideas behind the slope heuristics. The method relies on the two following points:

SH1: There exists a minimal penalty $\text{pen}_{\text{min}}(m)$ such that lighter penalties give rise to a selection of the most complex models, whereas higher penalties should select models with "reasonable" complexity.

SH2: An ideal penalty, that is a penalty leading to an oracle inequality, is about twice the minimal penalty.

In many contexts, a complexity measure C_m of the models is given. This complexity measure is typically the model dimension or the number of free parameters in parametric frameworks. Generally speaking, the penalty shape can be written as a function of C_m. When its definition is not obvious a priori, the complexity measure can be chosen as the penalty shape itself, as in Chapter 2 for instance. The penalty shape can also be guessed itself from the data, for example with resampling penalties.

For the two methods to apply the slope heuristics presented in further, it is required that:

(C1) The empirical contrast $\gamma_n(\hat{s}_m)$ decreases with the complexity C_m.

(C2) The penalty shape $\text{pen}_{\text{shape}}(\cdot)$ increases with the complexity C_m.

The two methods differ by the way the minimal penalty involved in point [SH1] is estimated. The first one is the so-called *dimension jump* method introduced in Birgé and Massart (2007). The second one consists of directly estimating the "slope" κ_{opt} in a data-driven fashion.

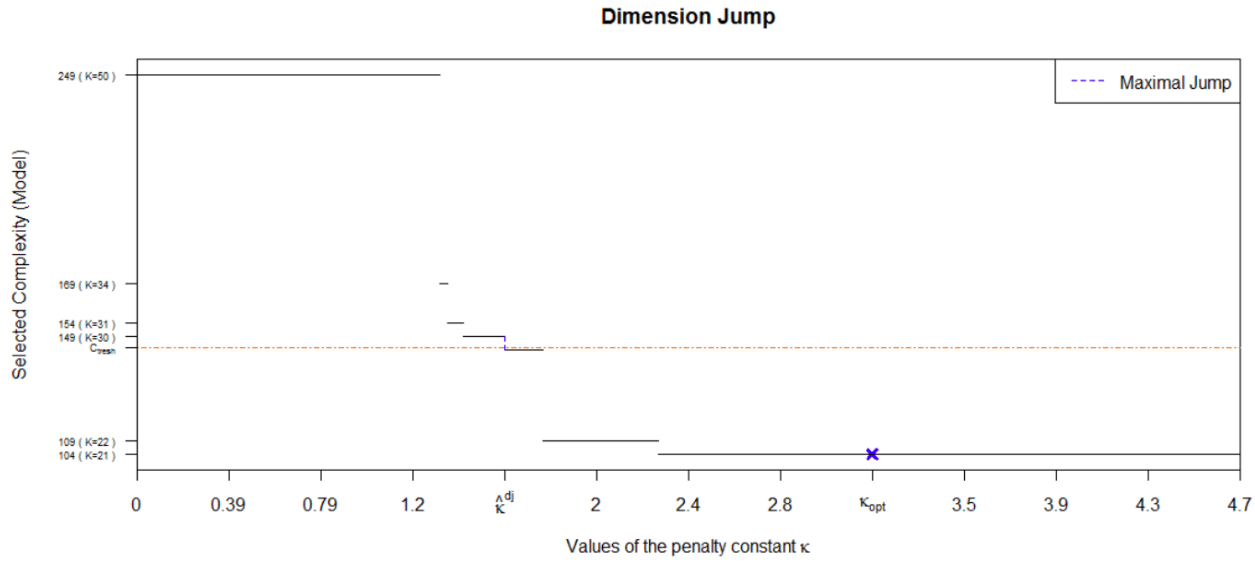


Figure 6.1: Graphical outputs provided by CAPUSHE. Representation of the nonincreasing and piecewise constant function $\kappa \mapsto C_{m(\kappa)}$.

6.2 Dimension jump

Principle. The so-called *dimension jump* is a method for penalty calibration which takes advantage of [SH1] and [SH2] to efficiently determine the unknown penalty constant κ_{opt} in (6.3). Let $m(\kappa)$ be the model selected by the penalized criterion $m \mapsto \gamma_n(\hat{s}_m) + \kappa \, \text{pen}_{\text{shape}}(m)$. Under (C1) and (C2), $\kappa \mapsto C_{m(\kappa)}$ is a nonincreasing and piecewise constant function. According to the minimal penalty definition, it is expected that the selected model $m(\kappa)$ has a large complexity when $\kappa \, \text{pen}_{\text{shape}}(\cdot) < \text{pen}_{\text{min}}(\cdot)$ and a reasonably large complexity if $\kappa \, \text{pen}_{\text{shape}}(\cdot) > \text{pen}_{\text{min}}(\cdot)$. Thus, $\kappa \mapsto C_{m(\kappa)}$ should present an abrupt jump around a value $\hat{\kappa}$ (see Fig. 6.1). The penalty $\hat{\kappa} \, \text{pen}_{\text{shape}}(\cdot)$ is then expected to be close to the minimal penalty and according to [SH2], the penalty $2\hat{\kappa} \, \text{pen}_{\text{shape}}(\cdot)$ is expected to be an optimal penalty ($\kappa_{\text{opt}} \approx 2\hat{\kappa}$).

Dimension jump algorithm. The dimension jump algorithm proceeds in three steps:

1. Compute, for all $\kappa > 0$, $m(\kappa) \in \text{argmin}_{m \in \mathcal{M}} \left\{ \gamma_n(\hat{s}_m) + \kappa \, \text{pen}_{\text{shape}}(m) \right\}$;
2. Find $\hat{\kappa}$ such that $C_{m(\kappa)}$ is large if $\kappa < \hat{\kappa}$ and has a "reasonable" order otherwise;
3. Select $\hat{m} = m(2\hat{\kappa})$.

This algorithm makes this first step computationally tractable since it only requires at most $\text{card}(\mathcal{M}) - 1$ steps, and actually probably much less. This provides the location of jumps, namely an increasing sequence $(\kappa_i)_{0 \leqslant i \leqslant i_{\max}}$ with $\kappa_0 = 0$, $\kappa_{i_{\max}} = +\infty$, the number of jumps $i_{\max} \in \{1, \ldots, \text{card}(\mathcal{M}) - 1\}$, and the associated selected model sequence $(m_i)_{0 \leqslant i \leqslant i_{\max}}$ where $m_i = m(\kappa_i)$ for all κ in $[\kappa_i, \kappa_{i+1})$ and for all $i < i_{\max}$. For the second step, two different strategies are available in CAPUSHE:

- **Maximal jump.** This first method is the most popular. It consists of choosing the constant $\hat{\kappa}^{\text{dj}}$ corresponding to the greatest jump of complexity: $\hat{\kappa}^{\text{dj}} = \kappa_{i_{\text{dj}}}$, with $i_{\text{dj}} \in \underset{0 \leqslant i \leqslant i_{\max} - 1}{\text{argmax}} \left\{ C_{m_{i+1}} - C_{m_i} \right\}$.
- **Threshold complexity.** The second method, proposed by Arlot and Massart (2009), consists of choosing a threshold complexity C_{thresh} such that complexities smaller than C_{thresh} are reasonable but larger ones are not. Then the chosen constant $\hat{\kappa}^{\text{thresh}}$ is the smallest value of κ for which the corresponding penalty selects a complexity smaller than C_{thresh}: $\hat{\kappa}^{\text{thresh}} = \inf\{\kappa > 0 : C_{m(\kappa)} \leqslant C_{\text{thresh}}\}$.

6.3 Data-driven slope estimation method

Principle. This alternative method consists of directly estimating the constant κ_{opt} by the "slope" of the expected linear relation of $-\gamma_n(\hat{s}_m)$ with respect to the penalty shape values $\text{pen}_{\text{shape}}(m)$. Indeed, it can be checked that $-\gamma_n(\hat{s}_m)$ behave linearly with respect to $\text{pen}_{\text{shape}}(m)$ with a slope around $\frac{\kappa_{\text{opt}}}{2}$, as shown in the left graph of Figure 6.2. Finally, if $\hat{\kappa}$ denotes an estimation of the slope of the linear regression of $-\gamma_n(\hat{s}_m)$ on $\text{pen}_{\text{shape}}(m)$, the optimal penalty is estimated by $2\hat{\kappa}\,\text{pen}_{\text{shape}}(\cdot)$. The package CAPUSHE proposes solutions so as to make possible and reliable the application of the slope heuristics thanks to a stability study of the selected model.

Practice of the data-driven slope estimation method. The main issue about this method is how to choose a subset of points $(\text{pen}_{\text{shape}}(m), -\gamma_n(\hat{s}_m))$ corresponding to large values of $\text{pen}_{\text{shape}}(m)$ where the slope can be estimated. In practice, it is usually chosen at sight. The method proposed in CAPUSHE to answer this problem is based on the model selection stabilization. More precisely, the slope is sequentially estimated from the couples $(\text{pen}_{\text{shape}}(m), -\gamma_n(\hat{s}_m))$ where the couple with the smallest penalty shape value is removed at each step. The slope estimation in this area corresponds to an estimation of $\kappa_{\text{opt}}/2$ and thus the same model is selected. Denoting $\mathcal{P} = \{\text{pen}_{\text{shape}}(m),\ m \in \mathcal{M}\}$, the corresponding algorithm consists of four consecutive steps.

- **Step 1** If several models in the collection have the same penalty shape value, only the model having the smallest contrast value $\gamma_n(\hat{s}_m)$ is kept according to (6.2). To make easier the reading of this algorithm, the model indexation is not modified.
- **Step 2** For any $p \in \mathcal{P}$, the slope $\hat{\kappa}(p)$ of the linear regression on the couples of points $\left\{(\text{pen}_{\text{shape}}(m), -\gamma_n(\hat{s}_m));\ \text{pen}_{\text{shape}}(m) \geqslant p\right\}$ is computed using a robust regression method.
- **Step 3** For any $p \in \mathcal{P}$, the model fulfilling the following condition is selected:

$$\hat{m}(p) = \underset{m\in\mathcal{M}}{\text{argmin}}\{\gamma_n(\hat{s}_m) + 2\hat{\kappa}(p)\text{pen}_{\text{shape}}(m)\}.$$

  We obtain an increasing sequence of change-points $(p_i)_{1\leqslant i\leqslant I+1}$ such that

$$\forall 1 \leqslant i \leqslant I-1, \forall p \in \mathcal{P}, \quad \begin{cases} \hat{m}(p) = \hat{m}(p_i) & \iff \quad p \in [p_i, p_{i+1}[\\ \hat{m}(p) = \hat{m}(p_I) & \iff \quad p \in [p_I, p_{I+1}]. \end{cases}$$

  We observe a "plateau" sequence and compute the plateau sizes $(N_i)_{1\leqslant i\leqslant I}$ defined by

$$\forall\ 1 \leqslant i \leqslant I-1,\ N_i = \text{card}\{[p_i, p_{i+1}) \cap \mathcal{P}\} \text{and} \quad N_I = \text{card}\{[p_I, p_{I+1}] \cap \mathcal{P}\}.$$

- **Step 4** The model $\hat{m}(p_{\hat{\imath}})$ such that $\hat{\imath} = \max\left\{i \in \{1,\dots,I\};\ N_i > pct\sum_{l=1}^{I} N_l\right\}$ is selected (see hereafter for the choixe of the *pct* value). We also return the interval of slope values $[p_{\hat{\imath}}, p_{\hat{\imath}+1})$ and the proportion $N_{\hat{\imath}}/\sum_{l=1}^{I} N_l$. Graphically, this corresponds to selecting the "most to the right" plateau whose length is greater than the threshold (see the bottom-right graph in Figure 6.2).

This algorithm requires to tune the parameter *pct* at Step 4 in order to determine which plateau corresponds to a stabilization of the model selection. By default, *pct* is set to 15% in CAPUSHE. This choice may be reconsidered according to the application at hand, and particularly to the size of $\mathcal{M}$ and to whether it is expected that many too complex models have been involved in the study. The experiments proposed in Baudry et al. (2012) suggest that the *pct* deeply impacts the model selection only in situations for which no linear behavior can be observed and on the contrary the method is not much sensitive to *pct* in favorable situations. Remark that whatever the choice at this step, the reported actual proportion $N_{\hat{\imath}}/\sum_{l=1}^{I-1} N_l$ measures the stability of the method: the higher this value the more confidently the method can be applied.

For the successive slope estimations in Step 2, a robust regression with the bisquare weighting function (Huber, 1981) is advised in order to attenuate the influence of possible estimation errors of the sequence $(\hat{s}_m)_{m\in\mathcal{M}}$.

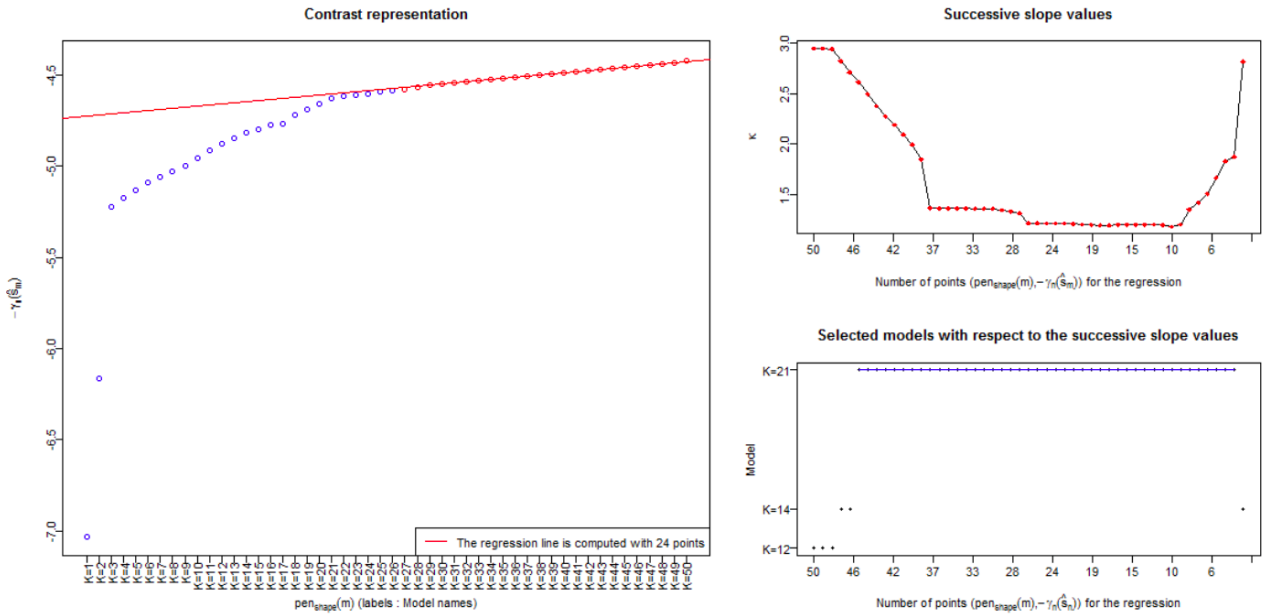


Figure 6.2: Graphical outputs provided by CAPUSHE. The left graph represents $-\gamma_n(\hat{s}_m)$ with respect to $\text{pen}_{\text{shape}}(m)$ to check the linear behavior assumption. The top-right (resp. bottom-right) graph gives the estimated slope (resp. the selected model) as a function of the number of couples $(\text{pen}_{\text{shape}}(m), -\gamma_n(\hat{s}_m))$ used for the linear regression. The last plateau for which the length $N_{\hat{i}}$ is greater than $pct \sum_{l=1}^{I} N_l$ is detected and the corresponding model $\hat{m}(p_{\hat{i}})$ is selected. The "Corresponding slope interval" given bottom right is the interval $[p_{\hat{i}}, p_{\hat{i}+1})$ leading to select $\hat{m}(p_{\hat{i}})$.

This method is based on a linear relation between $-\gamma_n(\hat{s}_m)$ and $\text{pen}_{\text{shape}}(m)$ for the largest values of the penalty shape. Non evidence of such linear relation should warn the user that the slope heuristics should probably not be applied. It should then be verified that complex enough models have been involved in the study and the penalty shape should be questioned. To help the user to validate the linear behavior assumption, some graphical tools are proposed in CAPUSHE.

Chapter 7

Feature selection for Random Forests

The work presented in this chapter has been carried out in the context of the industrial thesis of B. Gregorutti about airline safety (Gregorutti, 2015), in collaboration with the company Safety Lines[1]. Starting from an industrial problem, we have studied feature selection methods based on the permutation importance measure introduced by Breiman in the context of random forest algorithms (Gregorutti et al., 2013, 2014).

7.1 Random forests

Let us consider a variable of interest Y and a vector of random variables $\mathbf{X} = (X_1, \ldots, X_p)$. In the regression setting $s(x) = \mathbb{E}[Y|\mathbf{X} = x]$, a rule $\hat{s}$ for predicting Y is a measurable function taking its values in $\mathbb{R}$. The prediction error of $\hat{s}$ is then defined by $\mathcal{R}(\hat{s}) = \mathbb{E}\left[(\hat{s}(\mathbf{X}) - Y)^2\right]$.

Classification and regression trees, particularly CART algorithm due to Breiman et al. (1984), are competitive techniques for estimating s. Nevertheless, these algorithms are known to be unstable insofar as a small perturbation of the training sample may change radically the predictions. For this reason, Breiman (2001) introduced the random forests as a substantial improvement of the decision trees. It consists in aggregating a collection of such random trees, in the same way as the bagging method also proposed by Breiman (1996): the trees are built over n_{tree} bootstrap samples $\mathcal{D}_n^1, \ldots, \mathcal{D}_n^{n_{tree}}$ of the training data $\mathcal{D}_n$. Instead of CART algorithm, a subset of variables is randomly chosen for the splitting rule at each node. Each tree is then fully grown or until each node is pure. The trees are not pruned. The resulting learning rule is the aggregation of all of the tree-based estimators denoted by $\hat{s}_1, \ldots, \hat{s}_{n_{tree}}$. The aggregation is based on the average of the predictions.

7.2 Permutation importance measure and feature selection

The identification of the most relevant variables in high dimensional setting is a central issue in various applications. For linear regression, the Lasso method is widely used. Many variable selection procedures have also been proposed for non linear methods. In the context of random forests, it has been shown that the permutation importance measure introduced by Breiman, is an efficient tool for selecting variables (Díaz-Uriarte and Alvarez de Andrés, 2006; Genuer et al., 2010). Broadly speaking, a variable X_j can be considered as important for predicting Y if by breaking the link between X_j and Y the prediction error increases. To break the link between X_j and Y, Breiman proposes to randomly permute the observations of the X_j's. The empirical permutation importance measure can be formalized as follows: define a collection of out-of-bag samples $\{\bar{\mathcal{D}}_n^t = \mathcal{D}_n \backslash \mathcal{D}_n^t,\, t = 1, \ldots, n_{tree}\}$ which contains the observations not selected in the bootstrap subsets. Let $\{\bar{\mathcal{D}}_n^{tj},\, t = 1, \ldots, n_{tree}\}$ denotes a permuted out-of-bag samples by random permutations of the values of the j-th variable in each out-of-bag subsets. The empirical permutation importance of the variable X_j is defined by

$$\hat{I}(X_j) = \frac{1}{n_{tree}} \sum_{t=1}^{n_{tree}} \left[\hat{\mathcal{R}}(\hat{s}_t, \bar{\mathcal{D}}_n^{tj}) - \hat{\mathcal{R}}(\hat{s}_t, \bar{\mathcal{D}}_n^t) \right]. \tag{7.1}$$

[1] `http://www.safety-line.fr/en/`

where the empirical risk is defined for some $\bar{\mathcal{D}}$ by

$$\hat{\mathcal{R}}(\hat{s}, \bar{\mathcal{D}}) = \frac{1}{|\bar{\mathcal{D}}|} \sum_{i:(\mathbf{X}_i, Y_i) \in \bar{\mathcal{D}}} (Y_i - \hat{s}(\mathbf{X}_i))^2.$$

The quantity (7.1) is the empirical counterpart of the permutation importance measure $I(X_j)$, as formalized recently in Zhu et al. (2012). Let $\mathbf{X}_{(j)} = (X_1, \ldots, X'_j, \ldots, X_p)$ be the random vector such that X'_j is an independent replication of X_j which is also independent of Y and of all of the others predictors, the permutation importance measure is given by

$$I(X_j) = \mathbb{E}\left[\left(Y - s(\mathbf{X}_{(j)})\right)^2\right] - \mathbb{E}\left[(Y - s(\mathbf{X}))^2\right].$$

The permutation importance measure can be used to rank and to select the predictors. Nevertheless variable selection is a difficult issue especially when the predictors are highly correlated. In Gregorutti et al. (2013) we investigate deeper how the permutation importance measure depends on the correlation between the predictors. If $(\mathbf{X}, Y)$ is assumed to be a normal vector it is possible to specify the permutation importance measure:

Proposition 4. [Gregorutti et al. 2013] *Consider a Gaussian random vector*

$$(\mathbf{X}, Y) \sim \mathcal{N}_{p+1}\left(0, \begin{pmatrix} C & \boldsymbol{\tau} \\ \boldsymbol{\tau}^t & \sigma_y^2 \end{pmatrix}\right),$$

where $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_p)^t$ *with* $\tau_j = \mathcal{C}(X_j, Y)$, $\sigma_y^2 > 0$ *and* $C = [\mathcal{C}(X_j, X_k)]$ *is the non degenerated variance-covariance matrix of* $\mathbf{X}$. *For any* $j \in \{1, \ldots, p\}$, *let* $\alpha_j = [C^{-1}\boldsymbol{\tau}]_j$, *then*

$$I(X_j) = 2\alpha_j^2 \mathbb{V}(X_j) = 2\alpha_j \mathcal{C}(X_j, Y) - 2\alpha_j \sum_{k \neq j} \alpha_k \mathcal{C}(X_j, X_k).$$

The proposition confirms the impact of correlation on the importance measures, as noticed previously by Toloşi and Lengauer (2011) from experimental studies. More precisely, the proposition shows, first, that the independent variables may show higher importance values even if they are less informative than the correlated ones, and second, that the higher the number of correlated variables is, the faster the permutation importance of the variables decreases to zero.

For backward elimination strategies, these results also suggest that the permutation importance measure should be recomputed each time a variable is eliminated. More precisely, we follow the approach called Recursive Feature Elimination (RFE) inspired by Guyon et al. (2002) for SVM. It requires an updating of the permutation importance measures at each step of the algorithm. The RFE algorithm implemented in Gregorutti et al. (2013) can be summarized as follows:

1. Train a random forests
2. Compute the permutation importance measure
3. Eliminate the less relevant variable(s)
4. Repeat steps 1 to 3 until no further variables remain

By recomputing the permutation importance measure, we make sure that the ranking of the variables is consistent with their use in the current forest. We propose various simulation experiments in Gregorutti et al. (2013) to illustrate the efficiency of the RFE algorithm for selecting a small number of variables together with a good prediction error. Finally, this selection algorithm is aslo tested on the Landsat Satellite data from the UCI Machine Learning Repository[2].

[2] `https://archive.ics.uci.edu/ml/datasets/Statlog+(Landsat+Satellite)`

7.3 Grouped variable importance measure

In many situations, as in genetics studies, groups of variables can be clearly identified and it is of interest to select groups of variables rather than to select variables individually. In Gregorutti et al. (2014), we extend the variable importance measure for a group of variables $\mathbf{X}_J = (X_{j_1}, ..., X_{j_k})$.

For any $m \in \{1, \ldots, M\}$, let $\bar{\mathcal{D}}_n^{mJ}$ be a permuted version of $\bar{\mathcal{D}}_n^m$ obtained by randomly permuting the group $\mathbf{X}_J$ in each out-of-bag sample $\bar{\mathcal{D}}_n^{mj}$. Note that the same random permutation is used for each variable X_j of the group. By this way the (empirical) joint law of $\mathbf{X}_J$ is left unchanged by the permutation whereas the link between $\mathbf{X}_J$ and Y and the other predictors is broken. The empirical importance of $\mathbf{X}_J$ is defined by

$$\hat{\mathcal{I}}(\mathbf{X}_J) = \frac{1}{M} \sum_{m=1}^{M} \left[\hat{R}(\hat{f}_m, \bar{\mathcal{D}}_n^{mJ}) - \hat{R}(\hat{f}_m, \bar{\mathcal{D}}_n^m) \right]. \tag{7.2}$$

As for the importance of individual variables, $\hat{\mathcal{I}}(\mathbf{X}_J)$ is the empirical counterpart of a grouped variable importance measure that can be defined in a straightforward way.

We show that the grouped variable importance is equal to the sum of the individual importances when the variables of the group are independent, in the case of additive regression models. But of course this property is lost as soon as the variables in the group are correlated.

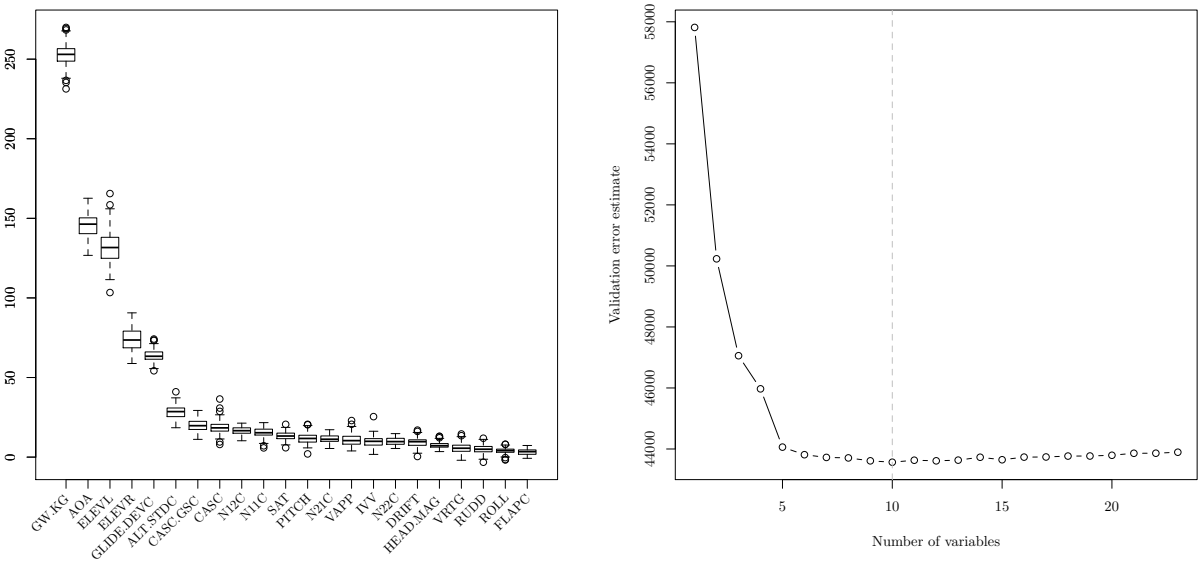
Application to multiple functional data analysis. The grouped variable importance can be used as a criterion for selecting features in the context of multiple functional regression. For instance, it can be fruitfully used for comparing the importances of wavelet coefficients in the context of functional predictors. Many groupings of wavelet coefficients can be proposed in this context. An important grouping strategy consists in grouping the wavelet coefficient related to a variable. As each group is associated to one variable, it is possible to obtain the importance of a given functional variable. Many other groupings could be proposed. As the wavelets are localized both in frequency and time, a group composed of all the wavelet coefficients of a given frequency level (for one or for all the variables) or a group composed by the wavelet coefficients associated to a given time t (to identify relevant time intervals) can be considered. One can also regroup two correlated variables. By computing the importances of such groups, one directly obtain the most important groups of coefficients for predicting the outcome. The RFE algorithm presented in the previous section is adapted for selecting groups of coefficients in a straightforward way. This backward grouped elimination approach produces a collection of nested subsets of groups.

7.4 A case study: variable selection for aviation safety

Airlines collect many informations during flights using flight data recorders. A large number of flight parameters are recorded as for instance the aircraft speed, the heading, the altitude or several warnings. In the context of the industrial thesis of B. Gregorutti about airline safety (Gregorutti, 2015), we have applied the RFE algorithm to a dataset of 1868 flights in order to identify the relevant variables to explain the risk of long landing. The evaluation of the risk the long landings is crucial for safety managers to avoid runway excursions and more generally to keep a high level of safety. Several risk factors have been identified among others the gross weight, the altitude or the angle of attack, see Fig. 7.1. Moreover some important time intervals have been detected for each functional variables.

7.5 Discussion and directions for future research

The idea behind Breiman's permutation importance measure is very general. It can be adapted to many problems by changing the risk function. Moreover, it can be computed for other learning methods than the random forests. In the context of a new Phd thesis in collaboration with the french electricity generator EDF, we are currently working on adapting the ideas presented in this chapter for estimating some functional of the quantile function of an outcome variable.



(a) (b)

Figure 7.1: Risk of long landing : (a) Boxplots of the grouped variable importance obtain for 100 runs of the selection algorithm, (b) MSE error versus the number of groups.

Publication list

Baudry, J.-P., Maugis, C., and Michel, B. (2012). Slope heuristics: overview and implementation. *Statistics and Computing*, 22(2):455–470.

Brault, V., Baudry, J.-P., Maugis, C., Michel, B., and Brault, M. V. (2011). *Package 'capushe'*.

Caillerie, C., Chazal, F., Dedecker, J., Michel, B., et al. (2011). Deconvolution for the wasserstein metric and geometric inference. *Electronic Journal of Statistics*, 5:1394–1423.

Caillerie, C. and Michel, B. (2011). Model selection for simplicial approximation. *Foundations of Computational Mathematics*, 11(6):707–731.

Chazal, F., Fasy, B. T., Lecci, F., Michel, B., Rinaldo, A., and Wasserman, L. (2014a). Robust topological inference: Distance to a measure and kernel distance. *ArXiv preprint 1412.7197*.

Chazal, F., Fasy, B. T., Lecci, F., Michel, B., Rinaldo, A., and Wasserman, L. (2015a). Subsampling methods for persistent homology. To appear in *Proceedings of the 32 st International Conference on Machine Learning (ICML-15)*.

Chazal, F., Glisse, M., Labruère, C., and Michel, B. (2014b). Convergence rates for persistence diagram estimation in topological data analysis. To appear in *Journal of Machine Learning Research*.

Chazal, F., Glisse, M., Labruère, C., and Michel, B. (2014c). Convergence rates for persistence diagram estimation in topological data analysis. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 163–171.

Chazal, F., Massart, P., and Michel, B. (2015b). Rates of convergence for robust geometric inference. ArXiv preprint 1505.07602.

Dedecker, J., Fischer, A., Michel, B., et al. (2015). Improved rates for wasserstein deconvolution with ordinary smooth error in dimension one. *Electronic Journal of Statistics*, 9:234–265.

Dedecker, J. and Michel, B. (2013). Minimax rates of convergence for wasserstein deconvolution with supersmooth errors in any dimension. *Journal of Multivariate Analysis*, 122:278–291.

Gaiffas, S. and Michel, B. (2014). Sparse bayesian unsupervised learning. ArXiv preprint 1401.8017.

Gregorutti, B., Michel, B., and Saint-Pierre, P. (2013). Correlation and variable importance in random forests. arXiv preprint 1310.5726.

Gregorutti, B., Michel, B., and Saint-Pierre, P. (2014). Grouped variable importance with random forests and application to multivariate functional data analysis. *To appear in Computational Statistics and Data Analysis*.

Maugis, C. and Michel, B. (2011a). Data-driven penalty calibration: a case study for gaussian mixture model selection. *ESAIM: Probability and Statistics*, 15:320–339.

Maugis, C. and Michel, B. (2011b). A non asymptotic penalized criterion for gaussian mixture model selection. *ESAIM: Probability and Statistics*, 15:41–68.

Maugis-Rabusseau, C. and Michel, B. (2013). Adaptive density estimation for clustering with gaussian mixtures. *ESAIM: Probability and Statistics*, 17:698–724.

Michel, B. (2008). *Modélisation de la production d'hydrocarbures dans un bassin pétrolier*. PhD thesis, Université Paris Sud-Paris XI.

Michel, B. (2011). Oil production: a probabilistic model of the hubbert curve. *Applied Stochastic Models in Business and Industry*, 27(4):434–449.

Bibliography

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory*, pages 267–281. Akadémiai Kiadó, Budapest.

Altun, K., Barshan, B., and Tunçel, O. (2010). Comparative study on classifying human activities with miniature inertial and magnetic sensors. *Pattern Recognition*, 43(10):3605–3620.

Amenta, N., Choi, S., Dey, T. K., and Leekha, N. (2000). A simple algorithm for homeomorphic surface reconstruction. In *Proceedings of the sixteenth annual symposium on Computational geometry*, pages 213–222. ACM.

Arias-Castro, E. and Verzelen, N. (2014). Detection and feature selection in sparse mixture models. *arXiv preprint arXiv:1405.1478*.

Arlot, S. (2009). Model selection by resampling penalization. *Electronic Journal of Statistics*, 3:557–624 (electronic).

Arlot, S. and Massart, P. (2009). Data-driven calibration of penalties for least-squares regression. *Journal of Machine Learning Research*, 10:245–279 (electronic).

Azizyan, M., Singh, A., and Wasserman, L. (2013). Minimax theory for high-dimensional gaussian mixtures with sparse mean separation. In *Advances in Neural Information Processing Systems*, pages 2139–2147.

Balakrishnan, S., Rinaldo, A., Sheehy, D., Singh, A., and Wasserman, L. A. (2012). Minimax rates for homology inference. *Journal of Machine Learning Research - Proceedings Track*, 22:64–72.

Banfield, J. and Raftery, A. (1993). Model-based gaussian and non-gaussian clustering. *Biometrics*, 49:803–821.

Barshan, B. and Yüksek, M. C. (2013). Recognizing daily and sports activities in two open source machine learning environments using body-worn sensor units. *The Computer Journal*, page bxt075.

Biau, G., Chazal, F., Cohen-Steiner, D., Devroye, L., and Rodriguez, C. (2011). A weighted k-nearest neighbor density estimate for geometric inference. *Electronic Journal of Statistics*, 5:204–237.

Birgé, L. (2005). A new lower bound for multiple hypothesis testing. *IEEE Trans. Inform. Theory.*, 51:1611–1615.

Birgé, L. and Massart, P. (2001). Gaussian model selection. *Journal of the European Mathematical Society*, 3:203–268.

Birgé, L. and Massart, P. (2007). Minimal penalties for Gaussian model selection. *Probability Theory Related Fields*, 138:33–73.

Blumberg, A. J., Gal, I., Mandell, M. A., and Pancia, M. (2014). Persistent homology for metric measure spaces, and robust statistics for hypothesis testing and confidence intervals. *Found. Comput. Math.*, pages 1–45.

Bobkov, S. and Ledoux, M. (2014). One-dimensional empirical measures, order statistics and Kantorovich transport distances. *Preprint*.

Bobrowski, O., Mukherjee, S., and Taylor, J. (2014). Topological consistency via kernel estimation. *arXiv preprint arXiv:1407.5272*.

Boissonnat, J.-D., Chazal, F., and Yvinec, M. (2015). Computational topology inference. Book to appear.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24:123–140.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). *Classification and regression trees*. Wadsworth Advanced Books and Software.

Bubenik, P. (2015). Statistical topological data analysis using persistence landscapes. *Journal of Machine Learning Research*, 16:77–102.

Bubenik, P. and Kim, P. T. (2007). A statistical approach to persistent homology. *Homology, Homotopy and Applications*, 9(2):337–362.

Buchet, M., Chazal, F., Dey, T. K., Fan, F., Oudot, S. Y., and Wang, Y. (2015a). Topological analysis of scalar fields with outliers. In *Proc. Sympos. on Computational Geometry*.

Buchet, M., Chazal, F., Oudot, S., and Sheehy, D. R. (2015b). Efficient and robust persistent homology for measures. In *Proceedings of the 26th ACM-SIAM symposium on Discrete algorithms. SIAM*. SIAM.

Burago, D., Burago, Y., and Ivanov, S. (2001). *A course in metric geometry*, volume 33. American Mathematical Society Providence.

Burnham, K. and Anderson, D. (2002). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. Springer-Verlag, New York, 2nd edition.

Butucea, C. and Tsybakov, A. B. (2008a). Sharp optimality in density deconvolution with dominating bias. I. *Theory Probab. Appl.*, 52:24–39.

Butucea, C. and Tsybakov, A. B. (2008b). Sharp optimality in density deconvolution with dominating bias. II. *Theory Probab. Appl.*, 52:237–249.

Cadre, B. (2006). Kernel estimation of density level sets. *Journal of multivariate analysis*, 97(4):999–1023.

Carlsson, G. (2009). Topology and data. *AMS Bulletin*, 46(2):255–308.

Carlsson, G., Jardine, R., Feichtner-Kozlov, D., Morozov, D., Chazal, F., de Silva, V., Fasy, B., Johnson, J., Kahle, M., Lerman, G., et al. (2012). Topological data analysis and machine learning theory.

Carroll, R. and Hall, P. (1988). Optimal rates of convergence for deconvolving a density. *J. Amer. Statist. Assoc.*, 83(404):1184–1186.

Chazal, F., Chen, D., Guibas, L., Jiang, X., and Sommer, C. (2011a). Data-driven trajectory smoothing. In *Proc. ACM SIGSPATIAL GIS*.

Chazal, F., Cohen-Steiner, D., Guibas, L. J., M'emoli, F., and Oudot, S. Y. (2009a). Gromov-hausdorff stable signatures for shapes using persistence. *Computer Graphics Forum (proc. SGP 2009)*, pages 1393–1403.

Chazal, F., Cohen-Steiner, D., and Lieutier, A. (2009b). Normal cone approximation and offset shape isotopy. *Computational Geometry*, 42(6):566–581.

Chazal, F., Cohen-Steiner, D., and Lieutier, A. (2009c). A sampling theory for compact sets in euclidean space. *Discrete & Computational Geometry*, 41(3):461–479.

Chazal, F., Cohen-Steiner, D., and Mérigot, Q. (2011b). Geometric inference for probability measures. *Foundations of Computational Mathematics*, 11(6):733–751.

Chazal, F., de Silva, V., Glisse, M., and Oudot, S. (2012). The structure and stability of persistence modules. *arXiv preprint arXiv:1207.3674*.

Chazal, F., De Silva, V., and Oudot, S. (2014d). Persistence stability for geometric complexes. *Geometriae Dedicata*, 173(1):193–214.

Chazal, F., Fasy, B. T., Lecci, F., Rinaldo, A., and Wasserman, L. (2014e). Stochastic convergence of persistence landscapes and silhouettes. In *Proc. 30th Annu. Sympos. Comput. Geom.*

Chazal, F., Guibas, L. J., Oudot, S. Y., and Skraba, P. (2013). Persistence-based clustering in riemannian manifolds. *Journal of the ACM (JACM)*, 60(6):41.

Chazal, F. and Lieutier, A. (2007). Stability and computation of topological invariants of solids in {\ Bbb R}ˆ n. *Discrete & Computational Geometry*, 37(4):601–617.

Chen, Y.-C., Genovese, C. R., and Wasserman, L. (2015). Density level sets: Asymptotics, inference, and visualization. *arXiv preprint arXiv:1504.05438*.

Cohen, S. and Le Pennec, E. (2013). Partition-based conditional density estimation. *ESAIM: Probability and Statistics*, 17:672–697.

Comte, F. and Lacour, C. (2011). Data driven density estimation in presence of unknown convolution operator. *J. Royal Stat. Soc., Ser B*, 73:601–627.

Cuevas, A. and Rodríguez-Casal, A. (2004). On boundary estimation. *Adv. in Appl. Probab.*, 36(2):340–354.

Cuevas, A. and Rodríguez-Casal, A. (2004). On boundary estimation. *Advances in Applied Probability*, pages 340–354.

De Silva, V. and Ghrist, R. (2007). Homological sensor networks. *Notices of the American mathematical society*, 54(1).

Delaigle, A., Hall, P., and Meister, A. (2008). On deconvolution with repeated measurements. *The Annals of Statistics*, pages 665–685.

Dereich, S., Scheutzow, M., and Schottstedt, R. (2013). Constructive quantization: Approximation by empirical measures. *Ann. Inst. H. Poincaré Probab. Statist.*, 49:1183–1203.

Devroye, L. and Wise, G. L. (1980). Detection of abnormal behavior via nonparametric estimation of the support. *SIAM J. Appl. Math.*, 38(3):480–488.

Díaz-Uriarte, R. and Alvarez de Andrés, S. (2006). Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*, 7:3.

Edelsbrunner, H. (1993). The union of balls and its dual shape. In *Proceedings of the ninth annual symposium on Computational geometry*, pages 218–231. ACM.

Edelsbrunner, H. and Harer, J. (2010). *Computational Topology: An Introduction*. AMS.

Edelsbrunner, H., Letscher, D., and Zomorodian, A. (2002). Topological persistence and simplification. *Discrete Comput. Geom.*, 28:511–533.

Fan, J. (1991). On the optimal rates of convergence for nonparametric deconvolution problems. *Ann. Statist.*, 19(3):1257–1272.

Fasy, B. T., Lecci, F., Rinaldo, A., Wasserman, L., Balakrishnan, S., and Singh, A. (2014). Confidence sets for persistence diagrams. *The Annals of Statistics*, 42(6):2301–2339.

Federer, H. (1959). Curvature measures. *Transactions of the American Mathematical Society*, pages 418–491.

Fournier, N. and Guillin, A. (2013). On the rate of convergence in wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, pages 1–32.

Genovese, C. R., Perone-Pacifico, M., Verdinelli, I., and Wasserman, L. (2012). Manifold estimation and singular deconvolution under hausdorff loss. *Ann. Statist.*, 40:941–963.

Genuer, R., Poggi, J.-M., and Tuleau-Malot, C. (2010). Variable selection using random forests. *Pattern Recognition Letters*, 31:2225–2236.

Gregorutti, B. (2015). *Forêts aléatoires et sélection de variables: analyse des données des enregistreurs de vol pour la sécurité aérienne*. PhD thesis, Université Pierre et Marie Curie-Paris VI.

Grove, K. (1993). Critical point theory for distance functions. In *Proc. Amer. Math. Soc. Symposia Pure Math*, volume 54, pages 357–385.

Guibas, L., Morozov, D., and Mérigot, Q. (2013). Witnessed k-distance. *Discrete Comput. Geom.*, 49:22–45.

Guyon, I., Weston, J., Barnhill, S., and Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine Learning*, 46(1-3):389–422.

Hatcher, A. (2001). *Algebraic Topology*. Cambridge Univ. Press.

Hatcher, A. (2002). *Algebraic topology*. Cambridge University.

Huber, P. J. (1981). *Robust Statistics*. Wiley.

Kasson, P. M., Zomorodian, A., Park, S., Singhal, N., Guibas, L. J., and Pande, V. S. (2007). Persistent voids: a new structural metric for membrane fusion. *Bioinformatics*, 23(14):1753–1759.

Kruijer, W., Rousseau, J., and van der Vaart, A. (2010). Adaptive Bayesian Density Estimation with Location-Scale Mixtures. *Electronic Journal of Statistics*, 4:1225–1257.

Lebarbier, E. (2005). Detecting multiple change-points in the mean of Gaussian process by model selection. *Signal Processing*, 85:717–736.

Lerasle, M. (2012). Optimal model selection in density estimation. In *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, volume 48, pages 884–908. Institut Henri Poincaré.

Lindsay, B. (1995). *Mixtures Models: Theory, Geometry and Applications*. IMS, Hayward, CA.

Massart, P. (2007). *Concentration Inequalities and Model Selection*, volume Lecture Notes in Mathematics 1896. Springer-Verlag.

McLachlan, G. and Peel, D. (2000). *Finite Mixture Models*. Wiley.

Meister, A. (2009). *Deconvolution problems in nonparametric statistics*. Lecture Notes in Statistics 193. Springer-Verlag.

Mérigot, Q. (2011). A multiscale approach to optimal transport. In *Computer Graphics Forum*, volume 30, pages 1583–1592. Wiley Online Library.

Meynet, C. and Maugis-Rabusseau, C. (2012). A sparse variable selection procedure in model-based clustering. Preprint.

Mileyko, Y., Mukherjee, S., and Harer, J. (2011). Probability measures on the space of persistence diagrams. *Inverse Problems*, 27(12):124007.

Munkres, J. R. (1984). *Elements of algebraic topology*, volume 2. Addison-Wesley Reading.

Neumann, M. H. and Hössjer, O. (1997). On the effect of estimating the error density in nonparametric deconvolution. *Journal of Nonparametric Statistics*, 7(4):307–330.

Niyogi, P., Smale, S., and Weinberger, S. (2008). Finding the homology of submanifolds with high confidence from random samples. *Discrete & Computational Geometry*, 39(1-3):419–441.

Phillips, J. M., Wang, B., and Zheng, Y. (2014). Geometric inference on kernel density estimates. *arXiv preprint 1307.7760*.

Polonik, W. (1995). Measuring mass concentrations and estimating density contour clusters-an excess mass approach. *The Annals of Statistics*, pages 855–881.

Rachev, S. and Rüschendorf, L. (1998). *Mass transportation problems*, volume II of *Probability and its Applications*. Springer-Verlag.

Saumard, A. (2010). Nonasymptotic quasi-optimality of aic and the slope heuristics in maximum likelihood estimation of density using histogram models.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6:461–464.

Singh, A., Scott, C., and Nowak, R. (2009). Adaptive Hausdorff estimation of density level sets. *Ann. Statist.*, 37(5B):2760–2782.

Singh, G., Mémoli, F., and Carlsson, G. E. (2007). Topological methods for the analysis of high dimensional data sets and 3d object recognition. In *SPBG*, pages 91–100. Citeseer.

Singh, G., Memoli, F., Ishkhanov, T., Sapiro, G., Carlsson, G., and Ringach, D. L. (2008). Topological analysis of population activity in visual cortex. *Journal of vision*, 8(8).

Tang, X. Y. (1994). Effect of dimension in multivariate deconvolution problems. Purdue University, Technical Report 94-8.

Toloşi, L. and Lengauer, T. (2011). Classification with correlated features: unreliability of feature ranking and solutions. *Bioinformatics*, 27:1986–1994.

Tsybakov, A. B. (2009). *Introduction to nonparametric estimation*, volume 11. Springer.

Tsybakov, A. B. et al. (1997). On nonparametric estimation of density level sets. *The Annals of Statistics*, 25(3):948–969.

Verzelen, N. (2010). Data-driven neighborhood selection of a gaussian field. *Computational Statistics & Data Analysis*, 54(5):1355–1371.

Villani, C. (2008). *Optimal Transport: Old and New*. Grundlehren Der Mathematischen Wissenschaften. Springer-Verlag.

von Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416.

Wasserman, L. (2014). Statistical inference for functional summaries of persistent homology. Topological Data Analysis Workshop 2014, Samsi.

Zhu, R., Zeng, D., and Kosorok, M. R. (2012). Reinforcement learning trees. Technical report, University of North Carolina.

Zomorodian, A. and Carlsson, G. (2005). Computing persistent homology. *Discrete Comput. Geom.*, 33(2):249–274.